

ANALYSIS OF DICHOTOMOUS QUESTIONS WITH IRT FOR DIAGNOSTIC TESTS IN STATISTICS COURSES

Andhita Dessy Wulansari
Institut Agama Islam Negeri Ponorogo
andhita@iainponorogo.ac.id

Abstract

The results of the question item analysis can be used as diagnostic information, whether students understand the concepts, have misconceptions, and need help understanding the concepts of the lecture's material. The diagnostic test here consists of 20 multiple-choice questions to identify the level of conceptual understanding of the primary statistics material in the introductory statistics course. In this study, researchers will conduct item analysis using Item Response Theory (IRT). This research aims to determine the questions' quality and the abilities of FTIK IAIN Ponorogo students. This research is evaluative research with a quantitative descriptive approach. The research results show that in terms of model suitability, based on the AIC, BIC, and log-likelihood criteria, for this case, the 1PL model is the most suitable compared to the 2PL and 3PL models. Based on the ICC 1PL curve, it can be seen that item number 9 is the most accessible item, and item number 4 is the most challenging item. To the right, the difficulty level of the question item is lower, and further to the left, the difficulty level of the question is higher. The estimated value of the person's ability (person) for each participant totaling 1000 people is symbolized by p_1 to p_{1000} . The test taker's ability (θ) varies from -3.45106849 to 3.38603802 .

Keywords : Dichotomous, IRT, Diagnostic, Statistics

Abstrak

Hasil analisis butir soal dapat digunakan sebagai informasi diagnostik, apakah mahasiswa paham konsep, miskonsepsi, dan tidak paham konsep atas materi yang telah diajarkan oleh dosen. Tes diagnostik disini terdiri dari 10 soal pilihan ganda untuk mengidentifikasi tingkat pemahaman konsep dari materi dasar-dasar statistika pada mata kuliah statistika dasar. Pada penelitian ini, peneliti akan melakukan analisis butir soal menggunakan Item Response Theory (IRT) dengan model 1PL, 2PL dan 3PL. Tujuan penelitian ini adalah untuk mengetahui bagaimana kualitas butir soal dan kemampuan dari mahasiswa FTIK IAIN Ponorogo dengan model yang paling cocok tersebut. Penelitian ini merupakan penelitian evaluatif dengan pendekatan deskriptif kuantitatif. Hasil penelitian menunjukkan bahwa Dalam hal kecocokan model, berdasarkan kriteria AIC, BIC dan log likelihood maka dapat diketahui bahwa untuk kasus ini, model 1PL adalah yang paling cocok dibandingkan model 2PL dan 3PL. Berdasarkan kurva ICC 1PL dapat diketahui bahwa item nomor 9 adalah item yang paling mudah dan item nomor 4 adalah item paling sulit dan kekanan tingkat kesulitan item soal semakin rendah dan semakin ke kiri tingkat kesulitan soal semakin tinggi. Untuk nilai estimasi kemampuan person (orang) dari masing-masing peserta yang

berjumlah 1000 orang, yang disimbolkan dengan p_1 sampai p_{1000} . Kemampuan peserta tes (θ) bervariasi pada rentang $-3,45106849$ sampai dengan $3,38603802$.

Kata Kunci: Dikotomus, IRT, Diagnostik, Statistika

INTRODUCTION

The 2022 Program for International Student Assessment (PISA) research results were recently announced on December 5, 2023, and Indonesia is ranked 658 with a score of math (379), science (398), and reading (371), Indonesia's achievements at the international level are very worrying (Bilad, Zubaidah, & Prayogi, 2024; Hiltunen et al., 2023; Wijaya, Hidayat, Hermita, Alim, & Talib, 2024). One of the reasons why Indonesia's low achievement, especially in mathematics, is that our students need help understanding the concepts being taught (Agustyaningrum, Sari, Abadi, & Mahmudi, 2021; Kristidhika, Cendana, Felix-Otuorimuo, & Müller, 2020; Puspitasari & Mufit, 2021; Sudirman, Son, Rosyadi, & Fitriani, 2020). Understanding the fundamental concepts in this research are focused on statistics is very important (Casella & Berger, 2024; Gupta & Kapoor, 2020; Hahs-Vaughn & Lomax, 2020; Mertler, Vannatta, & LaVenita, 2021), especially for students of FTIK (Faculty of Tarbiyah and Teacher Training) IAIN (State Islamic Institute) Ponorogo so they can understand other concepts. Students who need help understanding basic concepts will need help understanding other related concepts (Jonassen & Carr, 2020).

A diagnostic assessment is needed to identify whether students understand the concepts or misconceptions and need help understanding the concepts of the material taught by the lecturer (Andariana, Zubaidah, Mahanal, & Suarsini, 2020; Putra & Hamidah, 2020; Suwono et al., 2021; Tumanggor, Kuswanto, & Ringo, 2020). In designing learning, lecturers must also think about the assessment of their learning. Assessment is an important component in learning (Maki, 2023). To diagnose students' initial abilities, test instruments are needed. In making test instruments, lecturers must know the quality of the question items. Analysis of the quality of the question items is very important so that pseudo-assessments do not occur which have the impact of not measuring students' true abilities. Preparing test instruments is a means of evaluating. Evaluation here functions to determine whether the learning objectives have been achieved and the quality of the questions prepared.

Analysis of question items helps improve the quality of question items and can be used as information. Analysis of questions in education can be done using two approaches, namely the classical and modern approaches (Shultz, Whitney, & Zickar, 2020; Widyaningsih, Yusuf, Prasetyo, & Istiyono, 2021). Classical question item analysis is the process of reviewing question items through information from students' answers to improve the quality of the question items using Classical test theory. Furthermore, modern question item analysis reviews items using Item Response Theory (IRT) or theory—answers to question items (Zsido, Teleki, Csokasi, Rozsa, & Bandi, 2020). Item Response Theory is

a theory that uses mathematical functions to link the chance of answering a question correctly and a student's ability.

Analysis of diagnostic test items in introductory statistics courses is very important for lecturers. This is because statistics is a subject that is considered difficult for the majority of FTIK students. The analysis carried out will help lecturers find the difficulties experienced by students in studying basic statistics, especially material on the basics of statistics. There are still lecturers who still need to analyze the question items. One reason is that there are too many calculations if you do the analysis manually and some lecturers are worried about leaks in the questions they create. This is the reason why the quality of the questions given to students is still low.

The results of interviews by researchers with statistics lecturers at FTIK IAIN Ponorogo show that 30% of lecturers have never carried out question item analysis, and the rest have already carried out question item analysis. Based on the results of these interviews, researchers feel it is necessary to determine the questions' quality and students' abilities. Based on the research background, this research aims to determine the quality of diagnostic test items and the distribution of students' basic statistical abilities.

METHOD

The research carried out is in the form of evaluation research. This research uses a quantitative descriptive method to analyze the quality of the tested diagnostic test items and the level of student ability in the IRT (Item Response Theory) model. Based on the number of parameter items, in general, the IRT model that is popularly used is 1 Parameter Logistic Model (1 Parameter Logistic Model/1PL Model), 2 Parameter Logistic Model (2 Parameter Logistic Model/2PL Model) and 3 Parameter Logistic Model (3 Parameter Logistic Model/3PL Model).

In the 1PL Model, the difficulty level of the questions symbolized as b is a point on the ability scale so that test takers have a 0.5 chance of answering correctly on a particular item. For example, if an item/question has $b = 2$, then the test taker's ability is required at least 2 to answer correctly with a chance of 0.5. The greater the b value, the greater the test taker's ability to answer correctly, with a chance of 0.5. The value of b is in the range - up to, but the value of b is in a suitable category if it is in the range -2 to 2 (Hambleton & Swaminathan, 2013). The following are the opportunities for the ability to answer correctly in the 1PL Model (Hambleton, 1991).

$$P(x_{ij} = 1 | \theta_i, b_j) = \frac{\exp(\theta_i - b_j)}{1 + \exp(\theta_i - b_j)} \quad (1)$$

In the 2PL Model, apart from involving the question difficulty level parameter (b) like the 1PL Model, it also involves the difference power parameter (a). The value of a describes the slope of the ICC at point b on a specific ability scale. Parameter functions to detect whether or not an item/item can differentiate a group in the aspect being measured

(according to the differences that exist within the group). The value of a ranges from $-$ to $+$, but the value of a can be categorized as good if it is 0 to 2 (Hambleton & Swaminathan, 2013). The following is the probability of being able to answer correctly in the 1PL Model if the parameter a is added which shows the direction of the slope in the normal ogive (Hambleton, 1991).

$$P(x_{ij} = 1 | \theta_i, b_j, a_j) = \frac{\exp(a_j(\theta_i - b_j))}{1 + \exp(a_j(\theta_i - b_j))} \quad (2)$$

The 3PL model, apart from involving the problem difficulty level parameters (b) and differential power (a), like the 2PL model, also involves pseudo guessing (c). Parameter c shows the opportunity for test takers with low ability to be able to answer correctly items/questions that have a level of difficulty above their ability. The value of c extends from the range 0 to 1 (Retnawati, 2014). The value of the parameter c can be categorized as good if the value of $c < 1/k$ (Hulin, Drasgow, & Parsons, 1983). The following are the correct answer ability opportunities for the 3PL Model (Hambleton, 1991).

$$P(x_{ij} = 1 | \theta_i, b_j, a_j, c_j) = c_j + (1 - c_j) \frac{\exp(a_j(\theta_i - b_j))}{1 + \exp(a_j(\theta_i - b_j))} \quad (3)$$

This research activity was conducted at FTIK (Faculty of Tarbiyah and Teacher Training) and IAIN (State Islamic Institute) Ponorogo. The data used in the following analysis of 1PL, 2PL, and 3PL items is the score of correct (1) - incorrect (0) answers on a diagnostic test on the basics of statistics material of 10 items, and the number of test participants is 1000 students. To estimate capabilities, this research used Maximum Likelihood Estimation (MLE). Model parameter measurements were carried out using the R program.

RESULTS AND DISCUSSION

1PL Model Coefficients

Analysis of the 1PL question items (1 parameter logistic model) will be carried out using the R program so that the 1PL model coefficients are obtained as follows in figure 1.

Based on the output in figure 1, it can be seen that the item difficulty level is in the range of -1.762 to 0.167 , so the difficulty of item (b) can be categorized as medium to easy because many of the b values are close to -2 . Furthermore, for the value of a or discriminant power, the output results above show a constant value for items 1 to 10, namely 1.477 , because the model used in the analysis is a 1-parameter logistic model. The log-likelihood value is -5279.147 , so it can be concluded that the test participants data can be suitable using the 1PL model.

```

Dffclt Dscrmn
item1 -0.1521414 1.477385
item2 -1.3056666 1.477385
item3 -0.4456467 1.477385
item4 -1.7618188 1.477385
item5 -0.4858269 1.477385
item6 -0.5508525 1.477385
item7 -0.7269149 1.477385
item8 -0.4898669 1.477385
item9 0.1666072 1.477385
item10 -1.4943139 1.477385

Coefficients:
Dffclt.item1 Dffclt.item2 Dffclt.item3 Dffclt.item4 Dffclt.item5
-0.152 -1.306 -0.446 -1.762 -0.486
Dffclt.item6 Dffclt.item7 Dffclt.item8 Dffclt.item9 Dffclt.item10
-0.551 -0.727 -0.490 0.167 -1.494
Dscrmn
1.477
Log.Lik: -5279.147
    
```

Figure 1. 1PL Model Coefficients

Response Pattern of the 1PL Model

For the response pattern of FTIK IAIN Ponorogo students using the 1PL model, the output from the R program can be displayed as follows:

	item1	item2	item3	item4	item5	item6	item7	item8	item9	item10	Obs	Exp
1	0	0	0	0	0	0	0	0	0	0	5	11.565
2	0	0	0	0	0	0	0	0	0	1	5	5.597
3	0	0	0	0	0	0	0	0	1	0	1	0.481
4	0	0	0	0	0	0	0	0	1	1	2	0.406
5	0	0	0	0	0	0	0	1	0	0	4	1.269
6	0	0	0	0	0	0	0	1	0	1	1	1.071
7	0	0	0	0	0	0	1	0	0	1	2	1.520
8	0	0	0	0	0	0	1	1	0	1	1	0.467
9	0	0	0	0	0	1	0	0	0	1	2	1.172
10	0	0	0	0	0	1	0	1	0	0	1	0.266
11	0	0	0	0	0	1	0	1	0	1	2	0.360
12	0	0	0	0	1	0	0	0	0	1	1	1.064
13	0	0	0	1	0	0	0	0	0	0	4	8.310
14	0	0	0	1	0	0	0	0	0	1	6	7.011
15	0	0	0	1	0	0	0	0	1	0	1	0.603
16	0	0	0	1	0	0	0	0	1	1	4	0.816
17	0	0	0	1	0	0	0	1	0	0	2	1.590
18	0	0	0	1	0	0	0	1	0	1	3	2.153
19	0	0	0	1	0	0	0	1	1	0	3	0.185
20	0	0	0	1	0	0	0	1	1	1	1	0.387
21	0	0	0	1	0	0	1	0	0	0	2	2.256
22	0	0	0	1	0	0	1	0	0	1	3	3.056
23	0	0	0	1	0	0	1	0	1	1	1	0.549
24	0	0	0	1	0	0	1	1	0	0	1	0.693
25	0	0	0	1	0	1	0	0	0	0	1	1.740
26	0	0	0	1	0	1	0	0	1	1	1	0.423
27	0	0	0	1	0	1	0	1	0	0	1	0.594
28	0	0	0	1	0	1	0	1	0	1	1	1.117
29	0	0	0	1	0	1	0	1	1	1	1	0.302
30	0	0	0	1	0	1	1	1	0	0	1	0.359
31	0	0	0	1	0	1	1	1	0	1	1	1.132
32	0	0	0	1	0	1	1	1	1	1	1	0.459
33	0	0	0	1	1	0	0	0	0	0	1	1.580
34	0	0	0	1	1	0	0	0	0	1	3	2.140
35	0	0	0	1	1	0	0	0	1	1	1	0.385
36	0	0	0	1	1	0	0	1	0	1	1	1.015
37	0	0	0	1	1	0	0	1	1	1	1	0.275
38	0	0	0	1	1	0	1	0	0	1	1	1.440

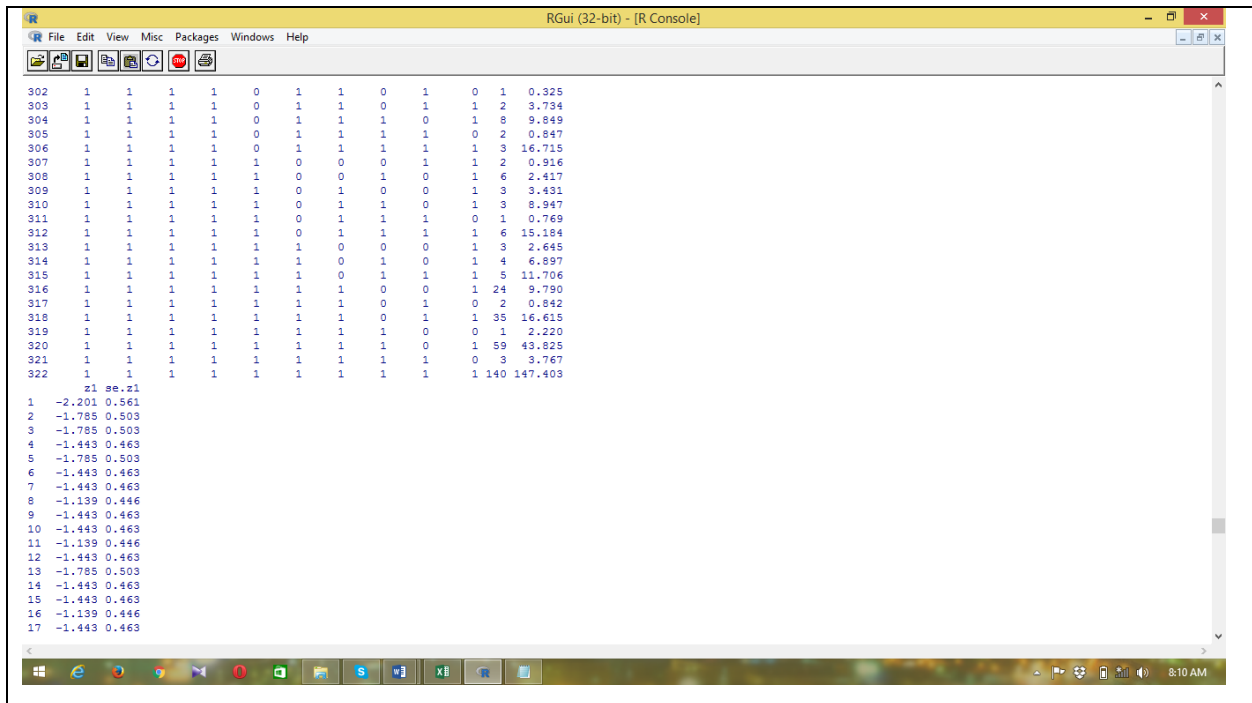


Figure 2 Response Patterns of 1PL Model Student Answers

Based on the output above, it can be seen that there were 322 response patterns created by 1000 test takers when working on these 10 items.

ICC Model IPL

The following is the item characteristic curve from the 1-parameter logistic model:

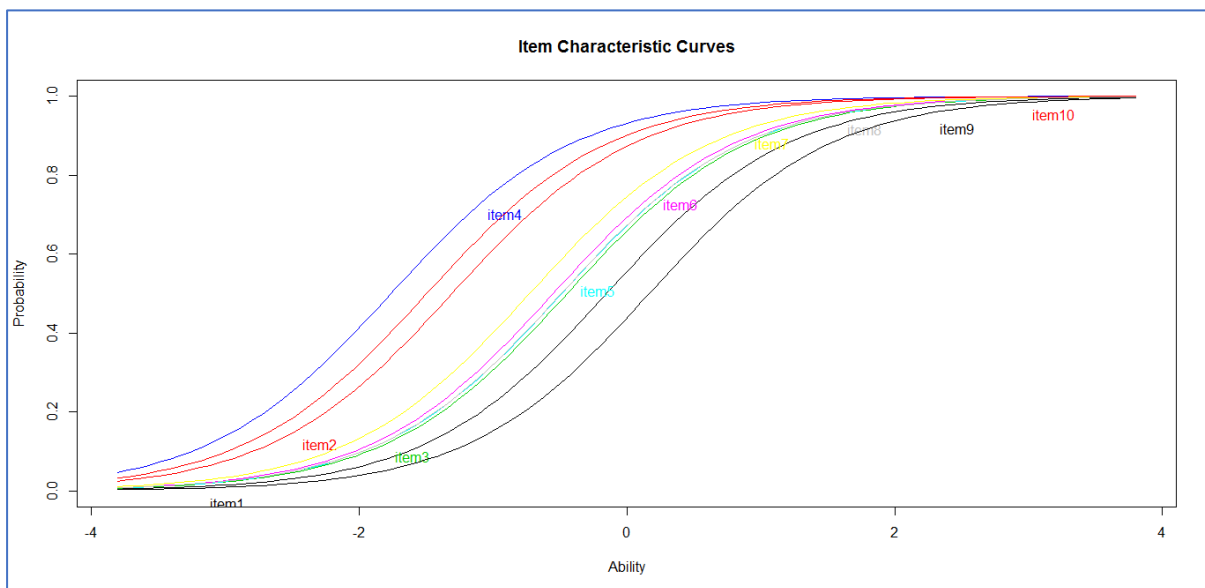


Figure 3. ICC of 1PL Model

Based on the curve shown in figure 3, it can be seen that item number 9 is the most accessible item, and item number 4 is the most challenging item. Explanation like this: if you look at it, a person with 0 ability when working on item 9 (rightmost item in the position of the curve), The probability of answering correctly is the smallest among the

other items, namely 0.4, whereas if the person works on item number 4 (the leftmost item in the position of the curve), then the probability of answering correctly is the largest among the other items, namely 0.9. So, in the ICC 1PL model above, the lower the level of difficulty of the question items, the lower the level of difficulty of the questions and the higher the level of difficulty.

2PL Model Coefficients

Analysis of the 2PL question items (2-parameter logistic model) will be carried out using the R program so that the 2PL model coefficients are obtained as follows:

	Dffc1t	Dscrmn
item1	-0.2567337	0.6895773
item2	-1.4041867	1.3032856
item3	-0.3934404	2.1844294
item4	-1.4042226	2.4705609
item5	-0.4071023	2.6533870
item6	-0.4835589	2.1562354
item7	-0.6258552	2.2135177
item8	-0.7138255	0.8308379
item9	0.2013040	1.0306464
item10	-1.3831443	1.7215991

Coefficients:		
	Dffc1t	Dscrmn
item1	-0.257	0.690
item2	-1.404	1.303
item3	-0.393	2.184
item4	-1.404	2.471
item5	-0.407	2.653
item6	-0.484	2.156
item7	-0.626	2.214
item8	-0.714	0.831
item9	0.201	1.031
item10	-1.383	1.722

Log.Lik: -5161.65

Figure 4. 2PL Model Coefficients

Based on the output above, it can be seen that the item difficulty level is in the range of -1.404 to 0.201 , so the difficulty of item (b) can be categorized between medium to easy because many of the b values are close to -2 . Furthermore, for the discriminant power value, the output results above show that the value is in the range of 0.690 to 2.653 , so the discriminant power (a) can be categorized as good because many of the values are close to 2 in the sense that the question items can differentiate test takers with high abilities—and low ability. The log-likelihood value is -5161.65 , so it can be concluded that the test participant data is also suitable when using the 2PL model.

Response Pattern of the 2PL Model

For the response pattern of FTIK IAIN Ponorogo students using the 2PL model, the output from the R program can be displayed as follow in figure 5.

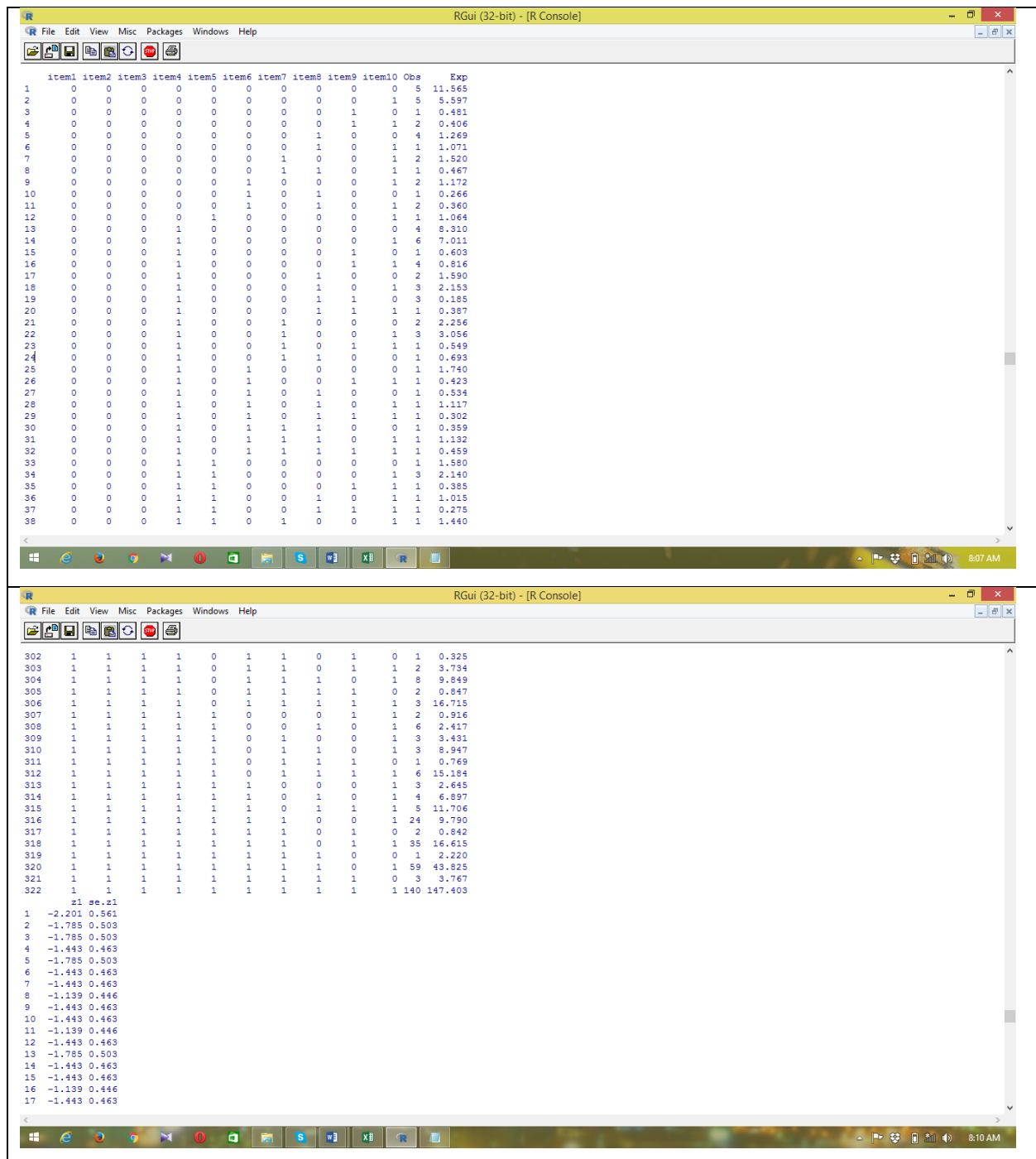


Figure 5. Response Patterns of 2PL Model Student Answers

Based on the output above, it can be seen that there were 322 response patterns created by 1000 test takers when working on these 10 items.

ICC Model 2PL

Figure 6 shows the item characteristic curve from the 2-parameter logistic model.

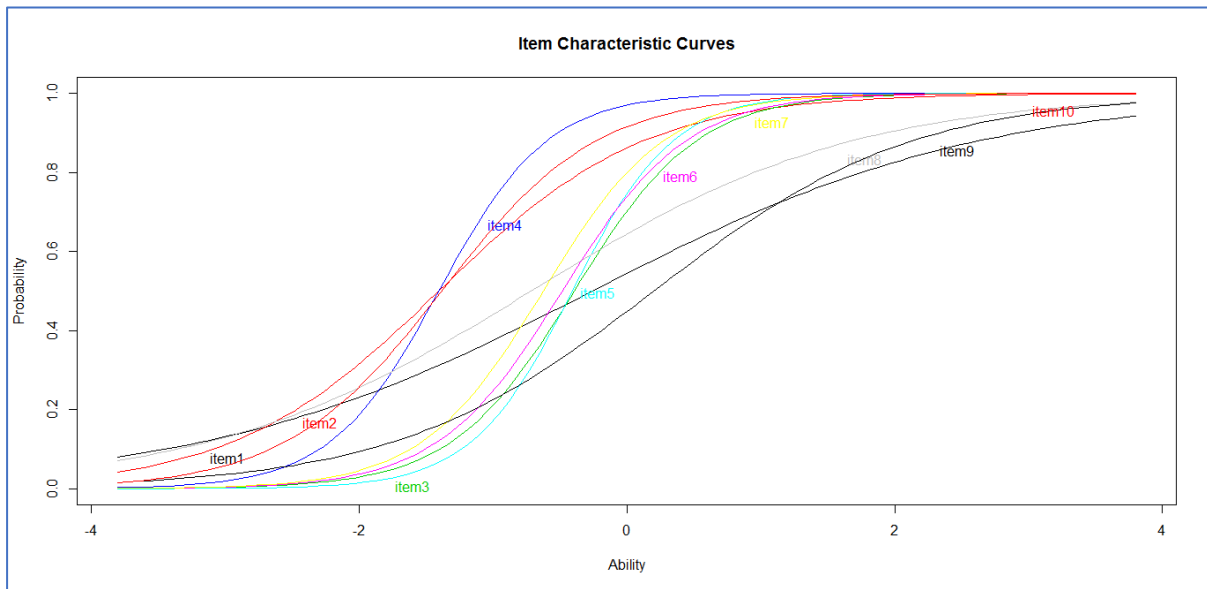


Figure 6. ICC of 1PL Model

Based on the curve image above, it can be seen that almost all question items have high (good) differentiation power. This can be seen from the graphs, which are almost all steep, although the level of steepness varies. For example, people with an ability of -2 to 0 when working on item 4 (high level of steepness) will have different probabilities of answering correctly; this shows that item 4 has good discrimination because it can differentiate the test taker's ability.

3PL Model Coefficients

Analysis of the 3PL question items (1 parameter logistic model) will be carried out using the R program so that the 3PL model coefficients are obtained as follows:

	Gussng	Dffclt	Dscrmn
item1	0.00318864	-0.2396865	0.6995005
item2	0.50309092	-0.3202501	2.3553396
item3	0.11148720	-0.1928668	2.8946112
item4	0.35618674	-0.9270460	4.1347782
item5	0.01818603	-0.3531554	2.7648583
item6	0.14485460	-0.2322023	2.9382587
item7	0.25234090	-0.2012189	4.1885541
item8	0.00671491	-0.6791174	0.8526141
item9	0.17180046	0.5801997	1.6536333
item10	0.41215538	-0.6940617	2.6259517

Coefficients:			
	Gussng	Dffclt	Dscrmn
item1	0.003	-0.240	0.700
item2	0.503	-0.320	2.355
item3	0.111	-0.193	2.895
item4	0.356	-0.927	4.135
item5	0.018	-0.353	2.765
item6	0.145	-0.232	2.938
item7	0.252	-0.201	4.189
item8	0.007	-0.679	0.853
item9	0.172	0.580	1.654
item10	0.412	-0.694	2.626

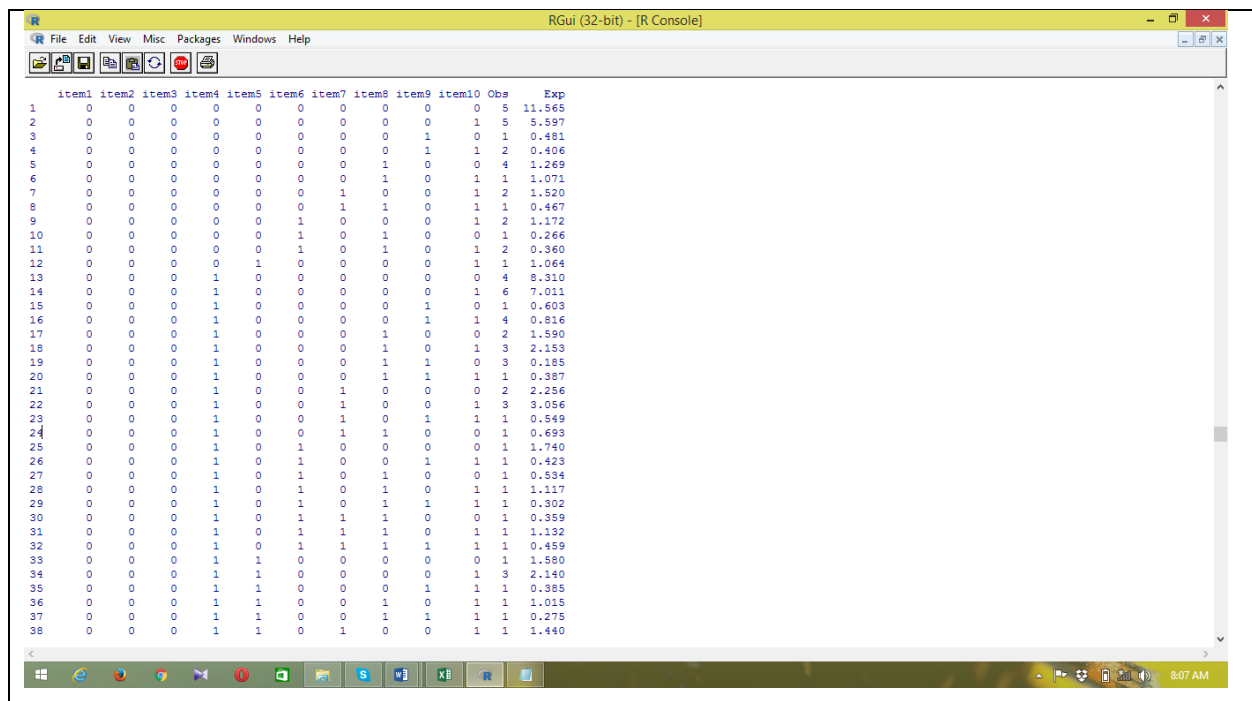
Log.Lik: -5139.395

Figure 7. 1PL Model Coefficients

Based on the output above, it can be seen that the item difficulty level is in the range of -0.927 to 0.580 , so the difficulty of item (b) can be categorized as medium to easy because many of the b values are close to -2 . Furthermore, for the discriminant power value, the output results above show that the value is in the range of 0.700 to 4.189 , so the discriminant power (a) can be categorized as being on the easy side because many of the values are above 2 , so that the question items can differentiate between test takers' abilities and Good. Meanwhile, the pseudo-guessing value (pseudo guessing) is in the range of 0.003 to 0.503 , so pseudo-guessing (c) can be categorized as being on the high side because many of the c values are above 0.25 ($1/\text{number of answers}$). The log-likelihood value is -5193.395 , so it can be concluded that the test participant data is also suitable when using the 3PL model.

Response Pattern of the 2PL Model

For the response pattern of FTIK IAIN Ponorogo students using the 2PL model, the output from the R program can be displayed as in figure 8.



	item1	item2	item3	item4	item5	item6	item7	item8	item9	item10	Obs	Exp
1	0	0	0	0	0	0	0	0	0	0	5	11.565
2	0	0	0	0	0	0	0	0	0	1	5	5.597
3	0	0	0	0	0	0	0	0	1	0	1	0.481
4	0	0	0	0	0	0	0	0	1	1	2	0.406
5	0	0	0	0	0	0	0	1	0	0	4	1.269
6	0	0	0	0	0	0	0	1	0	1	1	1.071
7	0	0	0	0	0	0	1	0	0	1	2	1.520
8	0	0	0	0	0	0	1	1	0	1	1	0.467
9	0	0	0	0	0	1	0	0	0	1	2	1.172
10	0	0	0	0	0	1	0	1	0	0	1	0.266
11	0	0	0	0	0	1	0	1	0	1	2	0.360
12	0	0	0	0	1	0	0	0	0	1	1	1.064
13	0	0	0	1	0	0	0	0	0	0	4	8.310
14	0	0	0	1	0	0	0	0	0	1	6	7.011
15	0	0	0	1	0	0	0	0	1	0	1	0.603
16	0	0	0	1	0	0	0	0	1	1	4	0.816
17	0	0	0	1	0	0	0	1	0	0	2	1.590
18	0	0	0	1	0	0	0	1	0	1	3	2.153
19	0	0	0	1	0	0	0	1	1	0	3	0.185
20	0	0	0	1	0	0	0	1	1	1	1	0.387
21	0	0	0	1	0	0	1	0	0	0	2	2.256
22	0	0	0	1	0	0	1	0	0	1	3	3.056
23	0	0	0	1	0	0	1	0	1	1	1	0.549
24	0	0	0	1	0	0	1	1	0	0	1	0.693
25	0	0	0	1	0	1	0	0	0	0	1	1.740
26	0	0	0	1	0	1	0	0	1	1	1	0.423
27	0	0	0	1	0	1	0	1	0	0	1	0.594
28	0	0	0	1	0	1	0	1	0	1	1	1.117
29	0	0	0	1	0	1	0	1	1	1	1	0.302
30	0	0	0	1	0	1	1	1	0	0	1	0.359
31	0	0	0	1	0	1	1	1	1	0	1	1.132
32	0	0	0	1	0	1	1	1	1	1	1	0.459
33	0	0	0	1	1	0	0	0	0	0	1	1.580
34	0	0	0	1	1	0	0	0	0	1	3	2.140
35	0	0	0	1	1	0	0	0	1	1	1	0.385
36	0	0	0	1	1	0	0	1	0	1	1	1.015
37	0	0	0	1	1	0	0	1	1	1	1	0.275
38	0	0	0	1	1	0	1	0	0	1	1	1.440

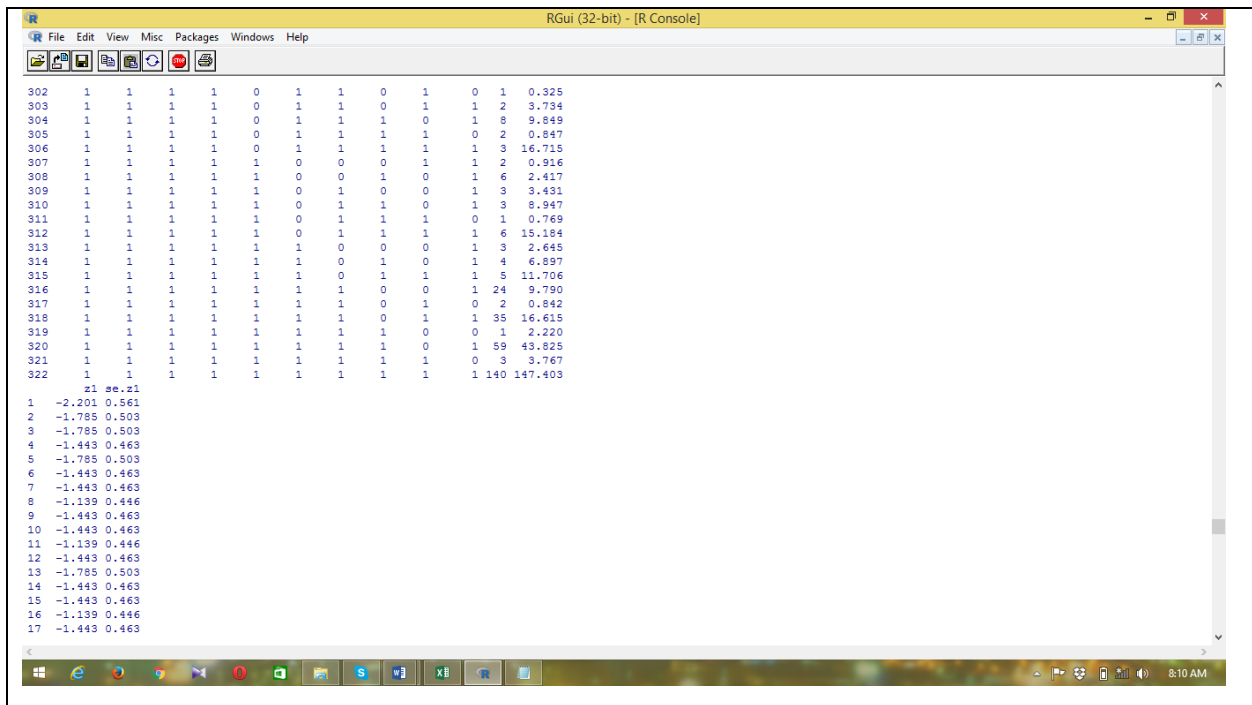


Figure 8. Response Patterns of 2PL Model Student Answers

Based on the output above, it can be seen that there were 322 response patterns created by 1000 test takers when working on these 10 items.

ICC Model 3PL

The following is the item characteristic curve from the 3-parameter logistic model:

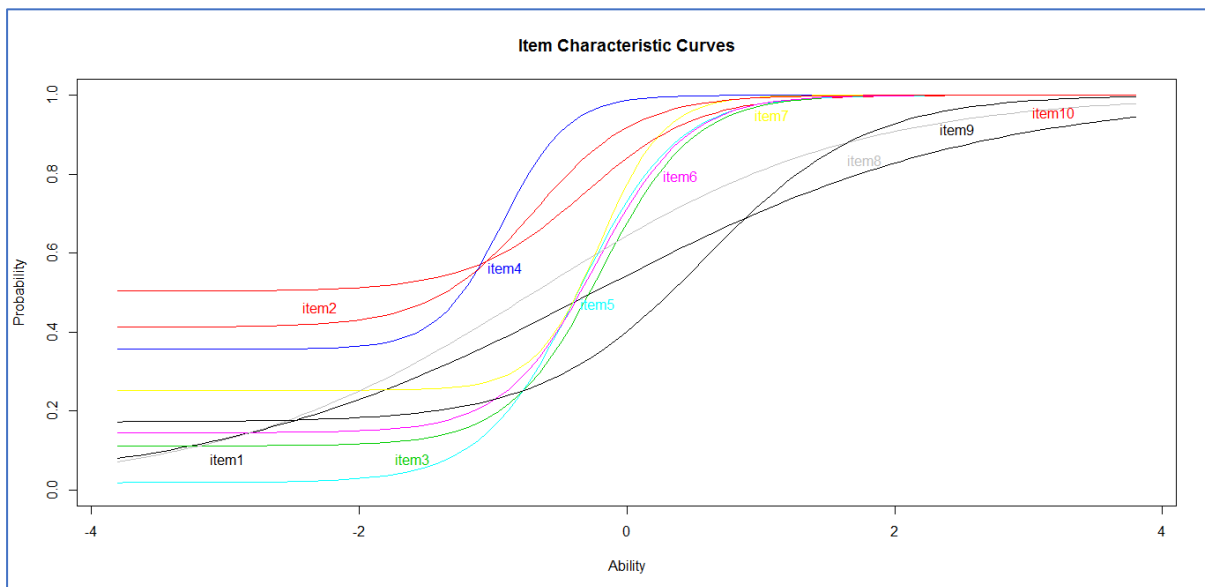


Figure 9. ICC of 3PL Model

Based on the curve shown in figure 9, it can be seen that there is guessing competition because people with different abilities have the same probability of answering correctly. This can be explained with the following example: for example, if a person with ability -2 to -4 answers question item number 2, the probability of answering the item

correctly is the same, namely 0.5 (shown by a straight line). So there is a guessing variable involved by the test taker when working on the question item.

1PL Model versus 2PL Model

Analysis of the differences between the 1PL and 2PL models was carried out using ANOVA so that the Likelihood Ratio Table for the 1PL model and the 2PL model was obtained as follows:

Likelihood Ratio Table						
	AIC	BIC	log.Lik	LRT	df	p.value
mml1	10580.29	10634.28	-5279.15			
mml2	10363.30	10461.45	-5161.65	234.99	9	<0.001

Figure 10. Likelihood Ratio Table for the 1PL model and the 2PL model

Based on the Likelihood Ratio Table, the p-value is <0.05 , so it can be concluded that different power factors can cause significant differences between the 1PL and 2PL models. In terms of model suitability, based on the AIC, BIC, and log-likelihood criteria, it can be seen that for this case, the 1PL model is more suitable than the 2PL model (it can be seen that the 1PL model criteria value is higher than the 2PL model).

1PL Model versus 3PL Model

Analysis of the differences between the 1PL and 3PL models was carried out using ANOVA so that the Likelihood Ratio Table for the 1PL model and the 3PL model was obtained as follows:

Likelihood Ratio Table						
	AIC	BIC	log.Lik	LRT	df	p.value
mml1	10580.29	10634.28	-5279.15			
mml3	10338.79	10486.02	-5139.39	279.51	19	<0.001

Figure 11. Likelihood Ratio Table for the 1PL model and the 3PL model

Based on the Likelihood Ratio Table, the p-value is <0.05 , so it can be concluded that different power factors and pseudo-guessing can cause significant differences between the 1PL and 3PL models. In terms of model suitability, based on the AIC, BIC, and log-likelihood criteria, it can be seen that for this case, the 1PL model is more suitable than the 3PL model (it can be seen that the 1PL model criteria value is higher than the 3PL model).

2PL Model versus 3PL Model

Analysis of the differences between the 1PL and 3PL models was carried out using ANOVA so that the Likelihood Ratio Table for the 1PL model and the 3PL model was obtained as follows:

Likelihood Ratio Table					
	AIC	BIC	log.Lik	LRT	df p.value
mml2	10363.30	10461.45	-5161.65		
mml3	10338.79	10486.02	-5139.39	44.51	10 <0.001

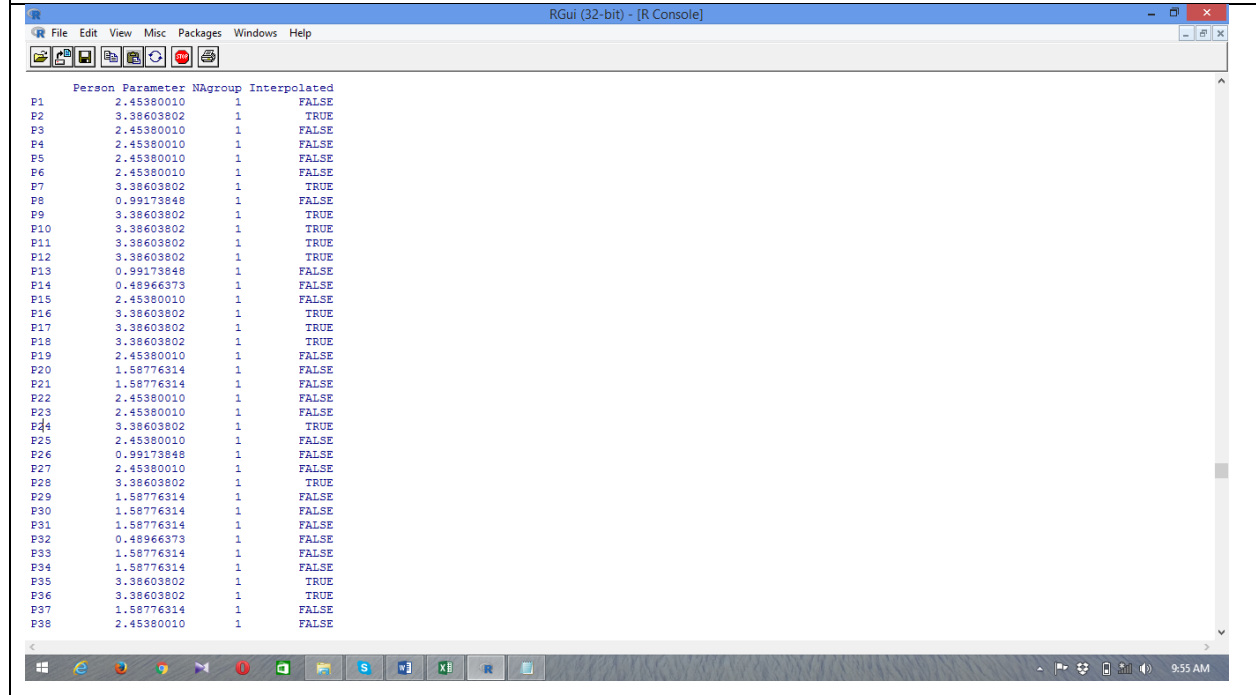
Figure 12. Likelihood Ratio Table for the 2PL model and the 3PL model

Based on the Likelihood Ratio Table, the p-value is <0.05 , so it can be concluded that pseudo-guessing can cause significant differences between the 2PL and 3PL models. In terms of model suitability, based on the AIC, BIC, and log-likelihood criteria, it can be seen that for this case, the 2PL model is more suitable than the 3PL model (it can be seen that the 2PL model criteria value is higher than the 3PL model).

Person Estimation Analysis

This person estimation analysis was also carried out using the R program, resulting in the following results:

beta item1	beta item2	beta item3	beta item4	beta item5	beta item6
0.84727092	-0.86007023	0.40808757	-1.51303983	0.34797630	0.25078281
beta item7	beta item8	beta item9	beta item10		
-0.01155253	0.34193945	1.32019883	-1.13159329		



Person	Parameter	NAgrou	Interpolated
P1	2.45380010	1	FALSE
P2	3.38603802	1	TRUE
P3	2.45380010	1	FALSE
P4	2.45380010	1	FALSE
P5	2.45380010	1	FALSE
P6	2.45380010	1	FALSE
P7	3.38603802	1	TRUE
P8	0.99173848	1	FALSE
P9	3.38603802	1	TRUE
P10	3.38603802	1	TRUE
P11	3.38603802	1	TRUE
P12	3.38603802	1	TRUE
P13	0.99173848	1	FALSE
P14	0.48966373	1	FALSE
P15	2.45380010	1	FALSE
P16	3.38603802	1	TRUE
P17	3.38603802	1	TRUE
P18	3.38603802	1	TRUE
P19	2.45380010	1	FALSE
P20	1.58776314	1	FALSE
P21	1.58776314	1	FALSE
P22	2.45380010	1	FALSE
P23	2.45380010	1	FALSE
P24	3.38603802	1	TRUE
P25	2.45380010	1	FALSE
P26	0.99173848	1	FALSE
P27	2.45380010	1	FALSE
P28	3.38603802	1	TRUE
P29	1.58776314	1	FALSE
P30	1.58776314	1	FALSE
P31	1.58776314	1	FALSE
P32	0.48966373	1	FALSE
P33	1.58776314	1	FALSE
P34	1.58776314	1	FALSE
P35	3.38603802	1	TRUE
P36	3.38603802	1	TRUE
P37	1.58776314	1	FALSE
P38	2.45380010	1	FALSE

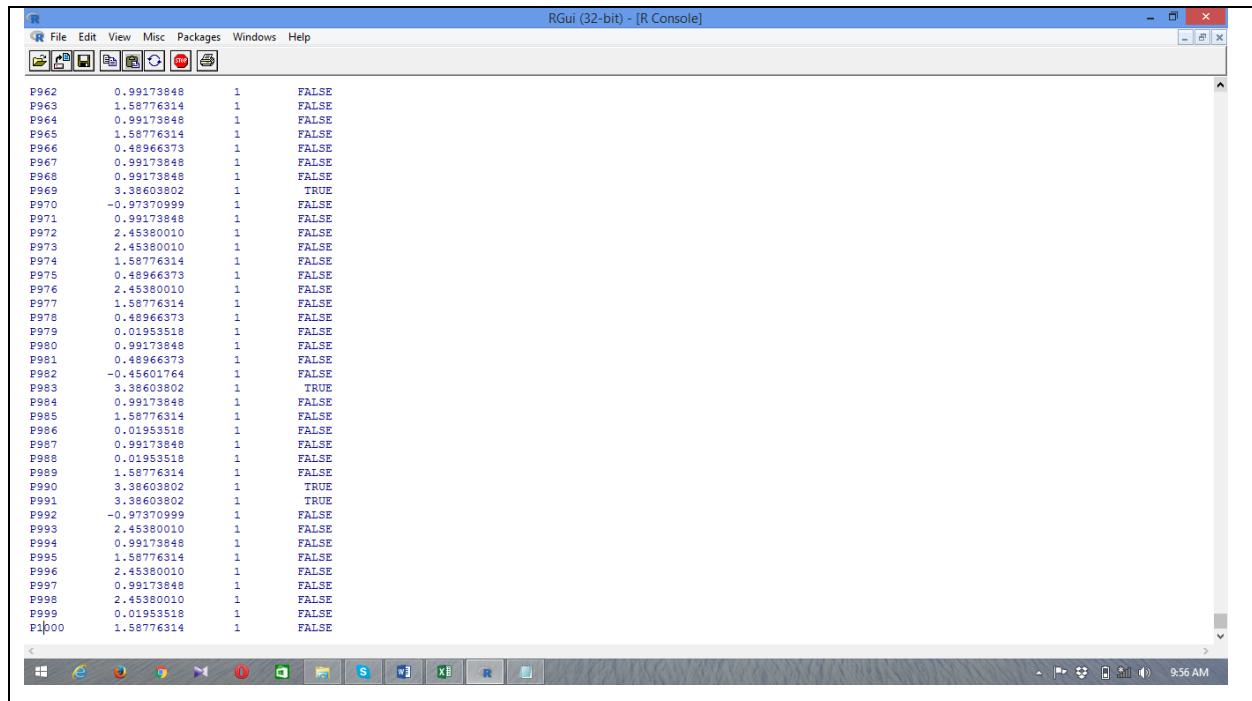


Figure 13. Person Estimation Analysis

CONCLUSION

The research results show that in terms of model suitability, based on the AIC, BIC, and log-likelihood criteria; for this case, the 1PL model is the most suitable compared to the 2PL and 3PL models. Based on the ICC 1PL curve, it can be seen that item number 9 is the most accessible item, and item number 4 is the most challenging item. To the right, the difficulty level of the question item is lower, and further to the left, the difficulty level of the question is higher. The estimated value of the person's ability (person) for each participant totaling 1000 people is symbolized by p_1 to p_{1000} . The test taker's ability (θ) varies from -3,45106849 to 3,38603802.

SUGESSTION

Based on the research "analysis of dichotomous questions with IRT for diagnostic tests in statistics courses," several suggestions were obtained for conducting further research related to the statistical abilities of FTIK IAIN Ponorogo students. Future research related to testing the initial statistical skills of FTIK IAIN Ponorogo students should use diagnostic tests that have been tested for the quality of the questions. Apart from that, in the future, the development of similar diagnostic test instruments but with variations in statistical material that differ in the quality of the tested items could enrich the choice of statistical ability test instruments that researchers in statistics can utilize. Therefore, the results of this research can be used as a reference in analyzing the quality of questions in the development of statistical ability diagnostic tests.

REFERENCES

- Agustyaningrum, N., Sari, R. N., et al. (2021). Dominant Factors That Cause Students' Difficulties in Learning Abstract Algebra: A Case Study at a University in Indonesia. *14*(1), 847-866.
- Andariana, A., Zubaidah, S., et al. (2020). Identification of biology students' misconceptions in human anatomy and physiology course through three-tier diagnostic test. *8*(3), 1071-1085.
- Bilad, M. R., Zubaidah, S., et al. (2024). Addressing the PISA 2022 Results: A Call for Reinvigorating Indonesia's Education System. *3*(1), 1-12.
- Casella, G., & Berger, R. (2024). *Statistical inference*: CRC Press.
- Gupta, S., & Kapoor, V. (2020). *Fundamentals of mathematical statistics*: Sultan Chand & Sons.
- Hahs-Vaughn, D. L., & Lomax, R. G. (2020). *An introduction to statistical concepts*: Routledge.
- Hambleton, R. K. (1991). *Fundamentals of item response theory*: Sage.
- Hambleton, R. K., & Swaminathan, H. (2013). *Item response theory: Principles and applications*: Springer Science & Business Media.
- Hiltunen, J., Ahonen, A., et al. (2023). PISA 2022 ensituloksia.
- Hulin, C. L., Drasgow, F., et al. (1983). *Item response theory: Application to psychological measurement*.
- Jonassen, D. H., & Carr, C. S. (2020). Mindtools: Affording multiple knowledge representations for learning. In *Computers as cognitive tools* (pp. 165-196): Routledge.
- Kristidhika, D. C., Cendana, W., et al. (2020). Contextual teaching and learning to improve conceptual understanding of primary students. *2*(2), 71-78.
- Maki, P. L. (2023). *Assessing for learning: Building a sustainable commitment across the institution*: Routledge.
- Mertler, C. A., Vannatta, R. A., et al. (2021). *Advanced and multivariate statistical methods: Practical application and interpretation*: Routledge.
- Puspitasari, R., & Mufit, F. (2021). *Conditions of learning physics and students' understanding of the concept of motion during the covid-19 pandemic*. Paper presented at the Journal of Physics: Conference Series.
- Putra, A., & Hamidah, I. (2020). *The development of five-tier diagnostic test to identify misconceptions and causes of students' misconceptions in waves and optics materials*. Paper presented at the Journal of Physics: Conference Series.
- Retnawati, H. J. N. M. (2014). Teori respons butir dan penerapannya [Item response theory and its application].
- Shultz, K. S., Whitney, D., et al. (2020). *Measurement theory in action: Case studies and exercises*: Routledge.
- Sudirman, S., Son, A. L., et al. (2020). Uncovering the Students' mathematical concept understanding ability: a based study of both students' cognitive styles dependent and independent field in overcoming the problem of 3D Geometry. *10*(1).
- Suwono, H., Prasetyo, T. I., et al. (2021). Cell Biology Diagnostic Test (CBD-Test) portrays pre-service teacher misconceptions about biology cell. *55*(1), 82-105.
- Tumanggor, A., Kuswanto, H., et al. (2020). *Using four-tier diagnostic test instruments to detect physics teacher candidates' misconceptions: Case of mechanical wave concepts*. Paper presented at the Journal of physics: conference series.

- Widyaningsih, S. W., Yusuf, I., et al. (2021). The Development of the HOTS Test of Physics Based on Modern Test Theory: Question Modeling through E-Learning of Moodle LMS. *14*(4), 51-68.
- Wijaya, T. T., Hidayat, W., et al. (2024). Exploring contributing factors to PISA 2022 mathematics achievement: Insights from Indonesian teachers. *13*(1), 139-156.
- Zsido, A. N., Teleki, S. A., et al. (2020). Development of the short version of the spielberger state—trait anxiety inventory. *291*, 113223.