Implementing Item Response Theory (IRT) Method in Quiz Assessment System

Viska Mutiawani, Al Misky Athaya, Kurnia Saputra, Muhammad Subianto

Informatics Department, Syiah Kuala University, Banda Aceh, Indonesia

Abstract **– A quiz is one of the methods to test students' abilities. Estimation of ability scores can be calculated by using the Classical Test Theory (CTT) and Item Response Theory (IRT) approaches. IRT is more sensitive to item characteristics that can estimate students' ability. This study implements IRT in a quiz assessment system, compares the results of the CTT and IRT, then analyses the functionality and usability of the built system. The system was tested by 50 respondents using 30 question items with 5 answer choices. The collected data is automatically assessed by the system using CTT and IRT calculations. The results showed differences in ranking between CTT and IRT methods. Using the IRT method, there were 14 respondents who experienced an increase in rank, 19 respondents experienced a decrease in rank, and the remaining 17 respondents did not experience a change in ranking. The functionality of built system was tested with the Blackbox Testing method and it has passed all the test functions. The usability of the system was also tested by involving 23 respondents using the System Usability Scale (SUS) questionnaire and the SUS score showed that the system can be accepted by users.**

Keywords **– Item Response Theory, Classical Test Theory, quiz assessment system, Black box testing, System Usability Scale.**

DOI: 10.18421/TEM111-26
https://doi.org/10.18421/TEM111-26

Corresponding author: Viska Mutiawani,
Informatics Department, Syiah Kuala University, Banda Aceh, Indonesia.
Email: viska.mw@unsyiah.ac.id

Received: 17 October 2021.
Revised: 20 January 2022.
Accepted: 27 January 2022.
Published: 28 February 2022.

1. Introduction

Education is something that everyone will do, either through formal or non-formal education processes. Education is the process of transferring knowledge from educators to students. This process can be held directly face to face (offline) or remotely (online). The online learning process (e-learning) has become highly common, either as an addition to the face-to-face process or as the main form of distance learning. Especially during the current COVID-19 pandemic which requires students to study from home, the online learning process is crucial.

The knowledge transfer process will often be assessed in the form of a hands-on test or a written test. A test is a method or procedure for measuring and assessing the educational process [1]. A test is an information-gathering tool that usually consists of a series of questions or exercises to measure students and measure the success of teaching programs [2]. Next, the tests or assessments will be discussed in this research focused on the written test that can measure test taker whether a test taker has mastered the subject being tested or not.

The form of questions on the written test can be in the form of multiple-choice questions, true-false, concise answers, equivalent answers, and essays. Written tests can be designed as an exercise for students so that after students answer, the correct answer will be provided. Written tests can also be designed as exams so that students answer at a predetermined time and date with a limited time span.

Previous research has produced a prototype of an offline quiz application that runs on a desktop computer so that the test can be done without the need for an internet connection [3]. The application prototype is an objective test system in the form of multiple-choice questions. The use of multiple-choice quiz forms has several advantages. Among the advantages are the material being tested can cover a wide scope, can measure abilities that vary from the simplest to the most complex, and a scoring system that is faster and easier.

The desktop-based offline quiz application that has been developed is still assessing in a simple way,

namely adding up the scores of the correct answers [3]. This score is considered as the grade of the student's ability and is known as the Classical Test Theory (CTT) approach. Charles Spearman put forward this classical test theory in 1904 and is widely applied in the education field [4]. The CTT is widely used because of its ease in calculating quiz scores. Measurement with the CTT approach shows a close relationship between the test item group and the test taker group [5]. If a test is done by a group of smart test-takers, the questions will be easy or the value of the difficulty index will be high. Then the same test is done by test takers who are less intelligent and then the questions will be difficult or the value of the difficulty index will be lower. So, the questions are inconsistent or changing reliant on the ability of the group of test-takers. This implies that the same test items must be answered by the same group of test-takers. If answered by a different group of test-takers, the item characteristics like the item difficulty and item discrimination can also change [6].

The limitation of the CTT in terms of item level information has made measurements with the Item Response Theory (IRT) approach to be widely studied and used. Based on the IRT theory, each question has a different weight value determined by the difficulty level of the question [7]. With this approach, it is hoped that the scores achieved by test takers can reflect their actual ability by taking into account other possibilities, for example the possibility of guessing the answers and the difficulty of the questions being worked on [8]. This IRT approach uses more assumptions than the CTT approach, but the IRT theory can provide more information.

The IRT approach focuses on information at the item level while the CTT focuses on information at the test level [9]. Therefore, this research will implement the IRT method on a quiz assessment system so that the resulting score can measure the ability of test-takers more accurately. This research will also compare the score obtained by the CTT method dan the score obtained by the IRT method. So, hopefully this research can benefit students and teachers/lecturers to get more information from assessment during the online learning process.

2. Literature Study

2.1. Item Response Theory (IRT)

Item Response Theory (IRT) is also known as Latent Trait Theory (LTT) or the Item Characteristic Curve (ICC) or Item Characteristic Function (ICF). The IRT method became widely known since the 1970s by measurement experts. In fact, the IRT was initially developed in the 1950s and 1960s by Frederic Lord and other psycho-metricians who intended to develop a method that is able to evaluate respondents without depending on the same items included in the test [10]. However, the attention to the IRT decreased until the late 1960s due to developments in the theory of the true score. When the True Score theory develops rapidly and attracts the attention of psychometric researchers, problems arise due to the weakness of the theory. Problems arose such as the lack of invariance of item parameters among the tested groups and the inability of classical test procedures to detect item bias. These things make the attention to the IRT increase again [11].

This IRT method actually aims to improve the weaknesses found in the CTT, namely the existence of group-dependent and item-dependent properties. In the CTT method, item difficulty, item discrimination, and item distractors depend on test-takers [12]. While in the IRT method, the value of the item difficulty level and other item characteristics will remain invariant to the group of test-takers [13]. Also, the measure of participants' ability will remain invariant to the test item group, it doesn't matter which item group they work on as long as the item group is able to be adequately done by the test taker [10].

The IRT mathematical model signifies that the subject's probability of answering the items accurately relies on the item characteristics and the subject's ability. This will make test takers with high or intelligent abilities to have a greater probability of answering accurately when compared to participants who have low abilities [6].

2.2. The IRT model

Item characteristic model depends on the mathematical form of the item characteristic function and the number of parameters used. Item parameters used in the IRT are item difficulty, item discrimination, and guessing probability (item distractors) [14]. There are 3 most common types of IRT models, namely one-parameter logistic model (1PLM or Rasch model), two-parameter logistic model (2PLM), and three-parameter logistics model (3PLM) [15]. The one-parameter logistic model (Rasch model) is a model used to analyze data that only focuses on the level of difficulty parameter [14]. The item characteristic curve for the one-parameter model is given by the following equation [6].

$$Pi(\theta) = \frac{e^{(\theta - b_i)}}{1+e^{(\theta - b_i)}} \qquad (1)$$

$Pi(\theta)$ = Probability of randomly selected participant with ability θ answers item i correctly.

e = Natural number whose value is close to 2.718.
θ = Participant's ability parameters.
b_i = Item i difficulty parameter

The level of difficulty parameter (b) describes the opportunity to correctly solve an item with a certain ability level. This value is originated from iteration or repetition of the participant's scores for each item being tested. Suppose Θ is the participant's ability and b is the level of difficulty, if Θ > b then it is assumed that the participant will be able to solve the item. If Θ < b, it is assumed that the participants will not be able to solve the question. If Θ/b = 1 then the participant has a 0.5 chance to answer the item correctly. This implies that the greater the value of the b parameter, the more difficult the item will be [6].

The IRT two-parameter logistic model (IRT 2PL) is a model used to analyze data that focuses on the level of difficulty and discriminatory parameters of questions [14]. The item characteristic curve for the two-parameter model is given by the following equation [6].

$$Pi(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1+e^{Da_i(\theta - b_i)}} \qquad (2)$$

$Pi(\theta)$ = Probability of randomly selected participant with ability θ answers item i correctly.
e = Natural number whose value is close to 2.718.
θ = Participant's ability parameters.
a_i = Item discrimination parameter.
D = Normal ogive function (D = 1.702).
b_i = Item i difficulty parameter

The discrimination parameter (a) is implemented to evaluate the ability of an item to differentiate between participants who have high abilities and participants who have low abilities. The value of discrimination is originated from iteration or repetition of the participant's scores for each item being tested. The usual range for item discrimination parameter is between 0 and 2 [6].

The IRT three-parameter logistic model (IRT 3PL) is a model used to analyze data that focuses on the parameters of difficulty level, discriminatory parameters of questions, and guessing probability (item distractor) [14]. The item characteristic curve for the 3-parameter model is given by the following equation [6].

$$Pi(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1+e^{Da_i(\theta - b_i)}} \qquad (3)$$

$Pi(\theta)$ = Probability of randomly selected participant with ability θ answers item i correctly.
e = Natural number whose value is close to 2.718.
θ = Participant's ability parameters.
a_i = Item discrimination parameter.
D = Normal ogive function (D = 1.702).
b_i = Item i difficulty parameter
c_i= Pseudo-chance-level parameter or guessing parameter.

The guessing parameter (c) describes the chance that participants with low ability to guess in answering the item correctly. The value of c is originated from iteration or repetition of the participant's scores for each item. In theory, the false guess values are in the range of 0 to 1 [16]. The value of the parameter c is calculated from the number of answer choices c = 1 / k, then k is the number of answer choices. For example, if the answer used in the study is in the form of 4 multiple choices, then the chances of guessing are 1/4 = 0.25 [8].

2.3. AdonisJS

Adonis JS is written in JavaScript programming language syntax. In 2019, the latest version of Adonis JS is 4.1. Node.js, which is the basis of Adonis JS, is software designed to develop web-based applications. If JavaScript is usually used to handle the client-side, then Node.js can also act as a programming language that runs on the server-side, such as PHP, Ruby, Perl, and so on. Node.js can run on Windows, Mac OS X, and Linux operating systems without the need for program code changes. Node.js has its own HTTP server library so it is possible to run a web server without using a web server program such as Apache or Nginx. Node.js was first created and introduced by Ryan Dahl, in 2009 so that JavaScript can be used as a server-side programming language, in the same class as PHP, ASP, Ruby, and so on [17].

Adonis JS has several advantages such as having a neat, simple, and easy-to-understand Syntax format. The Node.js on which Adonis JS is based is also capable of handling thousands of connections at the same time with minimum resource usage for each process which is called non-blocking. Besides that, Node.js also uses the JavaScript programming language as its base. Based on the research objectives previously described, Adonis JS will be a framework that builds a Quiz Assessment System to help understand how the IRT method is implemented.

2.4. Black Box Testing

Black Box Testing is testing software in terms of functional specifications without testing the design and program code [18]. Testing is intended to determine whether the functions, input, and output of the software conform with the required specifications. Black box testing is done by making test cases that try all the functions that are presented in the system whether they match the required specifications.

2.5. SUS questionnaire

There are several tests that software testers often use. Among them are Performance testing, System testing, Unit testing, Integration testing, Usability testing, Smoke testing, Stress testing, and User Acceptance test (UAT). Among usability testing, there is testing called System Usability Scale (SUS). SUS is a simple ten-question method that provides an overall view of the object to be tested [19]. The SUS was developed by John Brooke in 1986. It is a reliable, popular, effective, and inexpensive test scale. The SUS has 10 questions and 5 answer options. The answer choices consisted of strongly disagree to strongly agree. The final grade of the SUS has a minimum score of 0 and a maximum score of 100. The SUS has both positive and negative questions. Odd-numbered questions have a positive tone while even-numbered questions have a negative tone.

Usually, usability testing using the SUS uses the system's User Interface (UI) to get data from users. The user interacts with the system to determine whether a function has performed as expected and whether the UI makes the system easy to use. This test is often performed to get fast feedback in improving the interface and correcting errors in software components [19].

2.6. Previous Researches

Previous related research has been carried out by Mutiawani [3]. The output of that research is a desktop-based offline quiz application prototype. The quiz application can be utilized for students to answer quizzes offline. However, the process of updating data for example downloading and uploading quizzes still requires an internet connection. The offline quiz application still calculates the score by using the CTT method. Therefore, this research resumes previous research by adding the implementation of calculating scores using the IRT method.

Another related research was conducted by Liu [20] entitled " Design Flow of English Learning System Based on Item Response Theory". The result of this research is an English learning system based on the IRT which enables the system to provide personalized services for students and teachers.

Another related research conducted by Chen [21] developed an IRT-based e-learning application that considers the learning flow for students. The research was continued by Chen [22]. The result of the research is a fuzzy and IRT-based tutoring system application.

3. Research Methodology

The methodology of this research will follow three major steps which are literature study, quiz assessment system development, and IRT analysis. These steps can be seen in Fig. 1. The development process follows one of the Agile methodologies which is the eXtreme Programming (XP) method. Extreme Programming (XP) is a simple and fast software development method pioneered by Kent Beck, Ron Jeffries, and Ward Cunningham. XP can be used by a small to the medium-sized development team so it will be easier to accommodate any change in the development process very quickly [23]. The phases in the XP method will be explained in Figure 1.

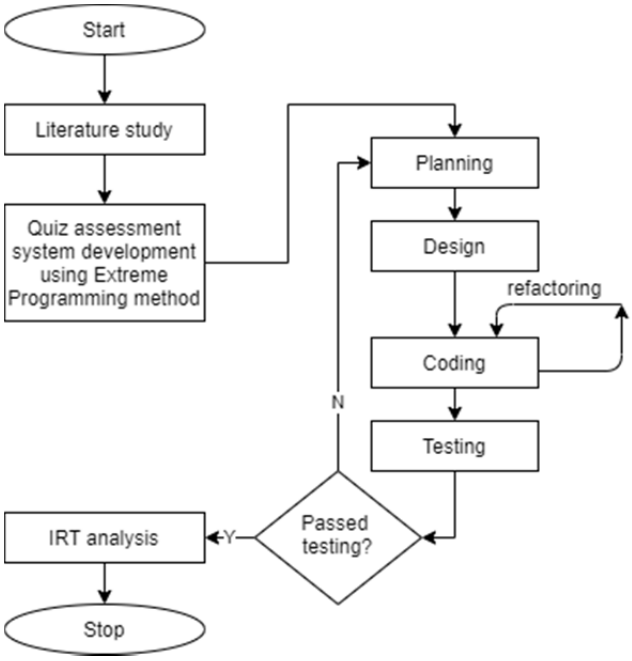

Figure 1. Research teps

3.1. Planning

This phase begins by collecting the requirements for an application that will be developed in the form of stories (user stories). These user stories will describe the required output, features, and functionality of the application.

3.2. Designing

This phase begins by collecting the requirements for an application that will be developed in the form of stories (user stories). These user stories will describe the required output, features, and functionality of the application.

3.3. Coding

After the stories were obtained and developed starting from the planning and design stage, the XP team did not immediately work on all the program codes, but first carried out a series of unit tests that ran each story. After a unit test is created, the developer will focus on what must be implemented in order to pass the unit test.

3.4. Testing

The first testing is conducted by using Black Box Testing to see the suitability of the input and the output produced by the system. Testing is focused on the features and functionality of the system. Testing is also carried out from the point of view of the user by using the System Usability Scale (SUS). This is done to find out whether the system has to meet user requirements or not. This test is conducted by distributing SUS questionnaires using Google Form.

4. Result and Discussion

4.1. Result

The first stage results in designing the database and the appearance of the quiz system. The functional design uses use case diagrams. The use case diagram describes the interactions that occur in the application so that it helps the developer to understand what functions are in the application and who can access these functions. In addition, use case diagrams are prepared to determine the needs and limitations of users in order to build a system that meets their requirements. The list of system requirements that were obtained in the previous stage resulted in a group of participants divided into 2 users, namely Admin and Quiz Participants. Admins can upload question banks, create quizzes, manage participants, and view quiz results. Meanwhile, quiz participants can do the quiz answer process. The list of system requirements that have been obtained previously is used to design the Entity-Relationship Diagram (ERD).

After the system design process is done, the next step is coding, which is the process of using a programming language in building pre-designed applications. This website-based online quiz application system was developed using Adonis.JS version 4.1 which is a framework from Node.JS. This framework uses the Javascript language as its foundation. For code editing, Visual Studio Code editor version 1.49.3 is the code editor. This application also uses MySQL as a database management system. Another software being used is R x64 v4.0.3 that was used as a backend software that calculates scores based on the IRT method.

First of all, the R software installation is performed on the server where the system is running. After that, an R script is created on the server which contains a list of the syntax used to perform the 3-parameter IRT model calculations. The system then runs a certain command that executes the syntax list in the R script that has been created. The results of the script execution are then translated by the system so that it is easy to read and finally displayed on the admin interface page. The pseudocode to run the R script from the quiz system can be seen in Figure 2.

```
Require: child process()
  arr = []
  result ← execute Rscript
  result → to STRING
  result ← SPLIT result
  for i = 1 < length of result do
    data ← SPLIT ← replace / to space in result
    arr ← data
  end for
  return arr
```

Figure 2. R script

The user interface depends on the role of the user. For quiz takers or participants, they are required to enter data in the form of a student ID number and PIN to log in to the system. The system will validate the data entered and direct the user to the dashboard page. The dashboard page contains a list of both available and past quizzes. Each quiz has its own detailed information such as the number of items, the length of time to answer, and the open and close date. Participants can only answer the quizzes that are available within the time frame listed on that page. When participants access outside the stated time, the button to answer the quiz will not be available.

Figure 3. is the interface page when participants answer the quiz. This page contains questions that will be answered by the participants along with the answer choices. The displayed question details are depended on the number selected by the participant on the right side of the page. Questions that have been answered are indicated by the color change of the number keypad from white to blue. This page also contains a number of time durations that will count down for the participants to answer the quiz. The page will close when the participant presses the submit button or when the available time duration has expired. When the available time duration has run out, an alert will appear notifying participants that the available time has been completely used and the system will direct participants to the dashboard page automatically.

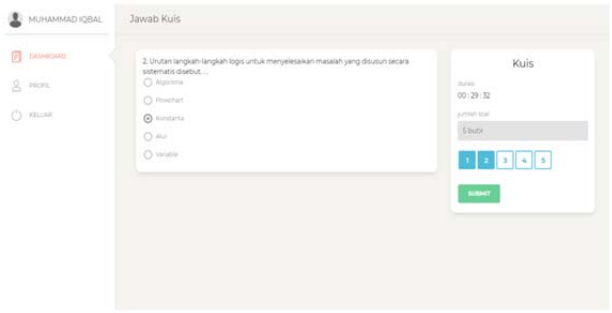

Figure 3. The quiz participant answering page

The user interface for admin also requires a login process. The admin is required to enter an ID and their password to enter the system. The system will validate the data entered and direct the user to the dashboard page. This page contains some information, namely, total participants, participants who have answered the quiz or not, total courses, total quizzes stored, and total available question banks. The data on this page will be updated automatically. To create a new quiz, the admin is required to insert the quiz details such as the course, the question bank, the name of the quiz, the open and close date, the quiz duration, and the number of quiz items. After the admin presses the button for the quiz, an alert will appear informing whether the quiz creation process is successful or not. Later, the created quiz, as well as question bank inserted, can be maintained by the admin.

Figure 4. is an interface page to see a list of scores from participants who have completed the quiz. Admins are required to select related courses and quizzes to see a list of quiz participant scores. This page contains detailed information on participant scores such as NPM, participant names, access times, submission times, quiz processing duration, and the number of correct points obtained from the total number of items.

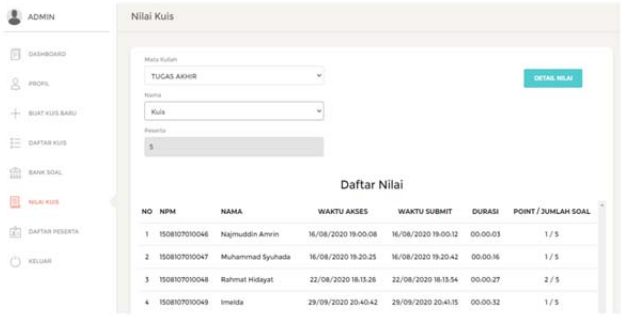

Figure 4. Quiz Score Page

Figure 5. is a page related to calculating scores using the IRT method. This page displays item characteristics for each item on the selected quiz. The item characteristics shown here are item difficulty, item discrimination, and guessing probability. This system uses IRT 3PL equation to calculate the correctness probability (P) for each item question. All values in Figure 5. are the data obtained from IRT calculations using R which is running on the server-side.

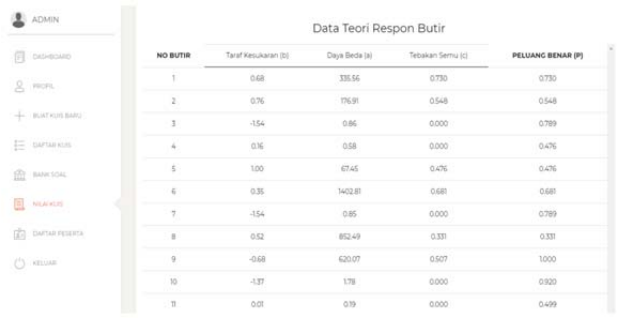

Figure 5. Item characteristics and correctness probability for IRT Score

This research compares the score obtained by using the CTT method and the score obtained by the IRT 3PL method. Figure 6. is the page that displays the ranking and score obtained by using the CTT method. While Figure 7. is an interface page to see the rankings of the quiz participants obtained using the IRT method calculation.

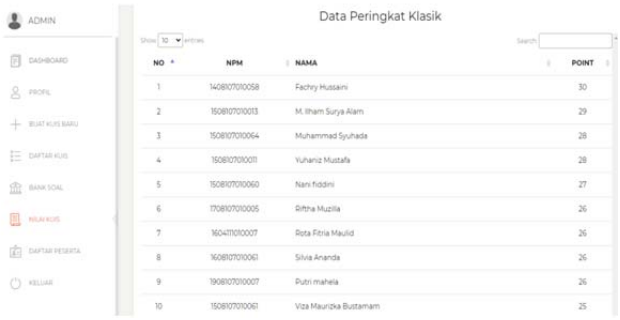

Figure 6. Participant rankings page using the CTT method

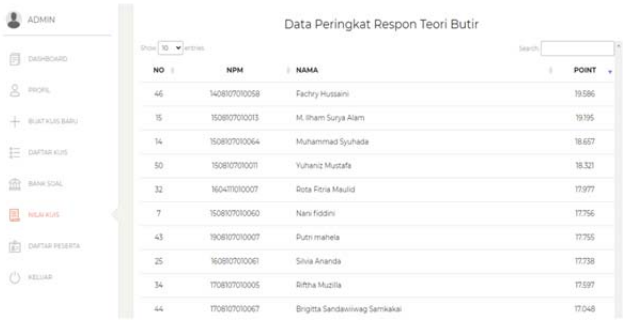

Figure 7. Participant rankings page using the IRT 3PL method

4.2. Discussion

The answers from participants who take the quiz are stored and then processed using an R script which returns all parameters needed to calculate the score using the IRT 3PL method. The IRT 3PL uses 3 parameters, namely item difficulty (b), item discrimination (a), and guessing (c). The parameters

obtained are used to get the correctness probability value for each item. The parameter values and the correctness probability obtained in this research can be seen in Table 1.

Table 1. Parameters and probability of each item

No.	*b*	*a*	*c*	*P*(θ)
1	0.68	335.56	0.730	0.730
2	0.76	176.91	0.548	0.548
3	-1.54	0.86	0.000	0.789
4	0.16	0.58	0.000	0.476
5	1.00	67.45	0.476	0.476
6	0.35	1402.81	0.681	0.681
7	-1.54	0.85	0.000	0.789
8	0.52	852.49	0.331	0.331
9	-0.68	620.07	0.507	1.000
10	-1.37	1.78	0.000	0.920
11	0.01	0.19	0.000	0.499
12	0.20	0.90	0.365	0.654
13	0.47	0.57	0.000	0.433
14	-1.10	0.89	0.000	0.726
15	-0.56	2.35	0.000	0.788
16	0.79	0.56	0.000	0.391
17	-1.10	1.20	0.000	0.788
18	-0.92	1.14	0.000	0.740
19	0.72	14.36	0.243	0.243
20	-0.68	567.97	0.188	1.000
21	-0.39	1.67	0.119	0.700
22	-0.68	507.27	0.000	1.000
23	-0.60	0.92	0.005	0.637
24	1.12	2.94	0.224	0.252
25	0.02	1.53	0.000	0.493
26	0.02	0.94	0.000	0.496
27	-0.98	2.57	0.000	0.926
28	-1.84	0.77	0.000	0.804
29	0.02	1109.51	0.586	0.586
30	-0.77	1.03	0.000	0.688

Based on Table 1., it can be seen that there are variations in the parameter values of item difficulty, item discrimination, and guessing probability on each item. According to [6], the item difficulty parameter (b) has ranged between -2.0 to +2.0 as a standard that determines items starting from easy to the difficult range. Items with a negative difficulty level indicate that these items are classified as easy (green in the table), and items with a positive difficulty level are classified as difficult (blue in the table). The level of difficulty in this study ranged from -1.84 to 1.11. As in Table 1., item number 28 has a difficulty level of -1.84 so it is classified as easy. The chances of participants being able to answer the item correctly are 0.804 or about 80%. This means that with a low level of difficulty, the chances of participants answering the item correctly are high and vice versa. Based on Table 1., there are 15 items questions that are classified as easy and 15 items that are classified as difficult.

As we can see from Table 1., the value of item discrimination (a) ranges from 0.57 to 1402.81. According to [6], a good discrimination value range between 0 and 2. Another opinion by [16] states that items that have a negative discrimination value are eligible to be eliminated. However, the data obtained in Table 1., shows that all items have a positive discrimination value which indicates that there are no items eligible to be eliminated. The negative discrimination value indicates that the items cannot accurately distinguish participants with high abilities from participants with low abilities.

According to [6], parameter c describes the opportunity for participants with low abilities to correctly answer an item that has a difficulty value that does not match the participant's ability. We can see that item number 1 has a high difficulty level and high false guessing. This means that the item has a high chance to be answered correctly in a predictable way despite the item's difficulty level is high. Items that fall under such criteria are eligible for elimination. The false guessing parameter has a value between 0 and 1. Items are classified as good if the value of c does not exceed c = 1 / k, where k is the number of answering choices [15]. There are 5 answer choices used in this study which makes the maximum limit of false guessing of 0.20. Based on this limit, items that have a value exceeding the maximum limit, namely items number 1,2,3,5,6,8,9,12,19,24, and 29 deserve to be eliminated.

The rankings of the participants are obtained in two different ways. Calculations using the CTT method apply the same weight to all tested items. This research determined the weight for each item is 1 point. These points are then accumulated for each participant who answers the questions correctly. Meanwhile, the calculation using the IRT method has a different weight or point for each item. This depends on the item difficulty level, item discrimination, and guessing probability for each question. These values are calculated and obtained from R x64 v4.0.3 software that has been installed on the server. These values are then used to obtain the Correctness Probability value for each item question. The use of the IRT method outcomes differ between one participant to another participant despite the number of items answered correctly is the same. This depends on the weight of the questions answered. It is possible for one participant to answer more questions with lower weight, whereas the other participants correctly answer more items that have a higher weight.

The total number of correct points for each participant obtained using both methods is then used to rank the participants. The participant with the highest number of points will be in the first rank while the participant with the lowest number of points is ranked last. The ranking data obtained from

the two methods were then compared. The result obtained is that there is a difference between the ranking using the CTT method and the ranking using the IRT 3PL method.

Table 2. Participant ranking comparison of CTT and IRT methods

Ranking	CTT method	IRT method
1	Respondent 1	Respondent 1
2	Respondent 2	Respondent 2
3	Respondent 3	Respondent 3
4	Respondent 4	Respondent 4
5	Respondent 5	Respondent 8
6	Respondent 6	Respondent 5
7	Respondent 7	Respondent 6
8	Respondent 8	Respondent 9
9	Respondent 9	Respondent 7
10	Respondent 10	Respondent 12
11	Respondent 11	Respondent 10
12	Respondent 12	Respondent 11
13	Respondent 13	Respondent 13
14	Respondent 14	Respondent 14
15	Respondent 15	Respondent 15
16	Respondent 16	Respondent 16
17	Respondent 17	Respondent 18
18	Respondent 18	Respondent 17
19	Respondent 19	Respondent 24
20	Respondent 20	Respondent 19
21	Respondent 21	Respondent 20
22	Respondent 22	Respondent 22
23	Respondent 23	Respondent 21
24	Respondent 24	Respondent 32
25	Respondent 25	Respondent 23
26	Respondent 26	Respondent 25
27	Respondent 27	Respondent 28
28	Respondent 28	Respondent 27
29	Respondent 29	Respondent 26
30	Respondent 30	Respondent 33
31	Respondent 31	Respondent 29
32	Respondent 32	Respondent 35
33	Respondent 33	Respondent 30
34	Respondent 34	Respondent 34
35	Respondent 35	Respondent 36
36	Respondent 36	Respondent 31
37	Respondent 37	Respondent 40
38	Respondent 38	Respondent 39
39	Respondent 39	Respondent 41
40	Respondent 40	Respondent 37
41	Respondent 41	Respondent 38
42	Respondent 42	Respondent 43
43	Respondent 43	Respondent 42
44	Respondent 44	Respondent 44
45	Respondent 45	Respondent 45
46	Respondent 46	Respondent 46
47	Respondent 47	Respondent 47
48	Respondent 48	Respondent 48
49	Respondent 49	Respondent 49
50	Respondent 50	Respondent 50

It can be seen in Table 2. that there is an increase and decrease in some of the participants' rankings. From 50 participants who took the quiz, there were 14 people experiencing a rank increase. The increase in participant ranking is indicated by the increase of position obtained using the IRT method compared to the position obtained using the CTT method. Table 2. also shows that 19 people experienced a decrease in ranking. The decrease in the ranking of participants is indicated by the decrease in the ranking position obtained using the IRT method when compared to the ranking position obtained using the CTT method. While the remaining 17 people did not experience a change in ranking.

Blackbox testing is used to test the functionality of the applications that have been built. This test was done by running the application with a predetermined scenario. The scenario contains a list of commands and features of the application being tested. The result is that all features are functioning well. The application was also tested for usability using the SUS questionnaire method. From 23 respondents, the SUS score was 79,7. This means that the application is good, acceptable and it has a grade C scale and adjective rating excellent as can be seen in Figure 8.

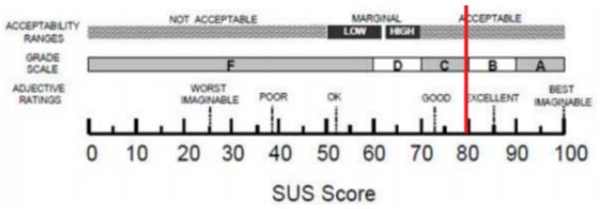

Figure 8. The meaning behind SUS score [24]

5. Conclusion and Future Work

Item Response Theory has been successfully implemented into an online quiz system in the form of multiple-choice with five answer choices. The rankings of some respondents are different between the ratings obtained from the CTT method and the ratings obtained from the IRT method. This proves that IRT makes each item has a different weight. This indicates that IRT was successfully implemented. The functionality of the quiz assessment system was tested using Black box testing and the result is all functions are functioning well. Testing the usability using the System Usability Scale (SUS) gets a score of 79.7. This means that the application is good, acceptable and it has a grade C scale and adjective rating excellent.

This application can be further developed. For example, adding the question file format accepted by the system, namely GIFT or AIKEN. The process of calculating the IRT that is carried out should also be able to run directly in the system without using third-party applications as helpers. For further research, it should also be done with more quiz participants and questions so that the differences in the scores obtained are more visible.

Acknowledgements

The authors would like to thank Syiah Kuala University for the chance of getting a research grant under the scheme "Penelitian Lektor, No. 270/UN11.2.1/PT.01.03/PNBP/ 2021". We also like to express our appreciation to the Informatics department staff for the support and motivation in doing this research.

References

[1]. Sudijono, A. (2001). *Pengantar evaluasi pendidikan*. Jakarta: Rajawali Press.

[2]. Arikunto, S. (2012). Dasar-dasar evaluasi pendidikan edisi 2. *Jakarta: Bumi Aksara*, *344*.

[3]. Mutiawani, V., Amrin, N., Saputra, K., & Yunardi, D. H. (2020, November). Developing a Desktop-based Offline Quiz Application. In *2020 IEEE Conference on e-Learning, e-Management and e-Services (IC3e)* (pp. 98-103). IEEE.

[4]. Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan rasch pada assessment pendidikan*. Trim komunikata.

[5]. Szabó, G. (2007). *Applying item response theory in language test item bank building* (Vol. 10). Peter Lang Pub Incorporated.

[6]. Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.

[7]. De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications.

[8]. Sudaryono, S. (2011). Implementasi Teori Responsi Butir (Item Response Theory) Pada Penilaian Hasil Belajar Akhir di Sekolah. *Jurnal Pendidikan dan Kebudayaan*, *17*(6), 719-732.

[9]. Abedalaziz, N., & Leng, C. H. (2018). The relationship between CTT and IRT approaches in Analyzing Item Characteristics. *MOJES: Malaysian Online Journal of Educational Sciences*, *1*(1), 64-70.

[10]. 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*.

[11]. Linden, W. J., & Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions. In *Handbook of modern item response theory* (pp. 1-28). Springer, New York, NY.

[12]. Hambleton, R. K., & Slater, S. C. (1997). Item response theory models and testing practices: Current international status and future directions. *European Journal of Psychological Assessment*, *13*(1), 21-28.

[13]. Hambleton, R. K., Robin, F., & Xing, D. (2000). Item response models for the analysis of educational and psychological test data. In *Handbook of applied multivariate statistics and mathematical modeling* (pp. 553-581). Academic Press.

[14]. Osterlind, S. J., & Wang, Z. (2017). Item response theory in measurement, assessment, and evaluation for higher education. In *Handbook on measurement, assessment, and evaluation in higher education* (pp. 191-200). Routledge.

[15]. Bichi, A. A., & Talib, R. (2018). Item Response Theory: An Introduction to Latent Trait Models to Test and Item Development. *International Journal of Evaluation and Research in Education*, *7*(2), 142-151.

[16]. Baker, F. B. (2001). *The basics of item response theory*. For full text: http://ericae. net/irt/baker..

[17]. Pitt, C. (2018). *The Definitive Guide to AdonisJs: Building Node. js Applications with JavaScript*. Apress.

[18]. Nidhra, S., & Dondeti, J. (2012). Black box and white box testing techniques-a literature review. *International Journal of Embedded Systems and Applications (IJESA)*, *2*(2), 29-50.

[19]. Brooke, J. (2013). SUS: a retrospective. *Journal of usability studies*, *8*(2), 29-40.

[20]. Liu, Y., & Zhao, X. (2017). Design Flow of English Learning System Based on Item Response Theory. *International Journal of Emerging Technologies in Learning (iJET)*, *12*(12), 91-102.

[21]. Chen, C. M., Lee, H. M., & Chen, Y. H. (2005). Personalized e-learning system using item response theory. *Computers & Education*, *44*(3), 237-255.

[22]. Chen, C. M., & Duh, L. J. (2008). Personalized web-based tutoring system based on fuzzy item response theory. *Expert systems with applications*, *34*(4), 2298-2315.

[23]. Pressman, R. S. (2019). *Software engineering: a practitioner's approach*. New York : McGraw-Hill Education.

[24]. Bangor, A., Kortum, P., & Miller, J. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of usability studies*, *4*(3), 114-123.