APPLYING THE ITEM RESPONSE THEORY WITH TWO-PARAMETER, THREE-PARAMETER MODELS IN THE EVALUATION OF MULTIPLE CHOICE TESTS

Nguyen Van Canh

Office of Quality Assurance, Dong Thap University, Vietnam

Corresponding author: nvcanh@dthu.edu.vn

Article history

Received: 26/11/2020; Received in revised form: 03/02/2021; Accepted: 06/4/2021

Abstract

The article presents the results of analyzing and evaluating multiple-choice items based on Item Response Theory (IRT) with two-parameter and three-parameter models through analysis results of data from R software (package ltm). Data in this study are the results of answering 50 multiple-choice items of 590 students who took the English 1 test organized at Dong Thap University in 2018. By evaluating each multiple-choice item based on their difficulty, discrimination parameters and guessing parameter according to the models, the study has identified good items to put into item bank, and point out items that are not really optimal, thus should continue to be considered before being put into use. The review and analysis of multiple-choice items based on both models help evaluate items more comprehensive and item selection more accurate. In addition, the research results show that if the evaluation of the test is only based on the subjective opinions of professional lecturers, not on the process of analyzing and evaluating based on IRT, the not good items could be introduced into the test without being detected.

Keywords: *Item Response Theory, multiple- choice items, two-parameter model, three-parameter model.*

ỨNG DỤNG LÝ THUYẾT ỨNG ĐÁP CÂU HỎI VỚI CÁC MÔ HÌNH 2 THAM SỐ, 3 THAM SỐ ĐÁNH GIÁ ĐỀ THI TRẮC NGHIỆM KHÁCH QUAN

Nguyễn Văn Cảnh

Phòng Đảm bảo chất lượng, Trường Đại học Đồng Tháp, Việt Nam

Tác giả liên hệ: nvcanh@dthu.edu.vn

Lịch sử bài báo

Ngày nhận: 26/11/2020; Ngày nhận chỉnh sửa: 03/02/2021; Ngày duyệt đăng: 06/4/2021

Tóm tắt

Bài viết trình bày kết quả phân tích, đánh giá đề thi trắc nghiệm khách quan dựa trên lý thuyết ứng đáp câu hỏi với các mô hình 2 tham số và 3 tham số thông qua kết quả phân tích dữ liệu từ phần mềm R (gói ltm). Dữ liệu được sử dụng trong nghiên cứu này là kết quả trả lời 50 câu hỏi trắc nghiệm khách quan của 590 sinh viên đối với đề thi Tiếng Anh 1 được sử dụng tại Trường Đại học Đồng Tháp năm 2018. Bằng việc đánh giá từng câu hỏi trắc nghiệm khách quan dựa trên các tham số độ khó, độ phân biệt và tham số đoán mò theo các mô hình, nghiên cứu đã chỉ ra những câu hỏi tốt có để đưa vào ngân hàng câu hỏi, đồng thời chỉ ra những câu hỏi chưa thật sự tối ưu cần phải tiếp tục được xem xét trước khi đưa vào sử dụng. Việc xem xét, phân tích các câu hỏi trắc nghiệm khách quan dựa trên cả hai mô hình giúp cho việc đánh giá câu hỏi được toàn diện hơn, đồng thời việc lựa chọn câu hỏi được chính xác hơn. Ngoài ra, kết quả nghiên cứu cho thấy nếu việc đánh giá đề thi chỉ dựa vào ý kiến chủ quan của giảng viên chuyên môn mà không trải qua quá trình phân tích, đánh giá dựa trên lý thuyết ứng đáp câu hỏi có thể không phát hiện được những câu hỏi chưa tốt và đưa vào các đề thi.

Từ khóa: *Câu hỏi trắc nghiệm lý thuyết ứng đáp câu hỏi, khách quan, mô hình 2 tham số, mô hình 3 tham số.*

DOI: https://doi.org/10.52714/dthu.10.4.2021.878

Cite: Nguyen, V. C. (2021). Applying the item response theory with two-parameter, three-parameter models in the evaluation of multiple choice tests. *Dong Thap University Journal of Science, 10*(4), 17-28. https://doi.org/10.52714/dthu.10.4.2021.878.

1. Introduction

Testing and assessment are important and indispensable activities in the teaching process, and it is a basis for adjusting teaching activities, which contributes to improving the quality of training. For assessment to be accurate, objective and for the learner's ability to be comprehensively evaluated, many universities are encouraging lecturers to participate in building exam item banks of many different types including multiple-choice items. Construction of item banks requires comprehensive expertise and evaluation of each item must be done by professional experts, and especially based on scientific theories of measurement in education, namely Classical Test Theory (CTT) and IRT. Although CTT is considered a meaningful theory, laying the foundation for the science of measurement in education, this theory has limitations. The major limitation of this theory is that separating test characteristics independently of the examinee's characteristics has not been done (Lam, 2011, p.76). However, with the introduction and strong development of IRT, the above limitations have been gradually overcome. Currently, the evaluation of multiple-choice item tests is often done by researchers using IRT through data statistics and analysis by specialized software. In this article, we apply R software to analyze, evaluate, and select multiple-choice items via IRT with two-parameter and three-parameter models. Using the *ltm* package, software R will calculate the difficulty, discrimination, and guessing parameters of each multiple-choice item. On that basis, the test editor can choose good items to put in item banks, and detect poor items that need removing or considering for before putting into use.

2. Literature Review

The science of educational measurement and evaluation in Vietnam was formed late and developed much slower than that in many countries in the world. However, this field has also been an interesting research theme by some educational managers, contributing to the development of this science in Vietnam. One of them is Duong Thieu Tong who had a research work on *Test and Measurement of Learning Achievement* in 1995. In his work, the author systematized the concepts of learning achievement measurement, principles of writing multiple-choice items and initially presented the analysis and evaluation of multiple-choice items based on CTT, and introduced a brief approach to IRT through the Rasch model. Additionally, there are research works by Lam Quang Thiep such as *Test and Application* in 2008, *Measurement in education Theory and application* in 2011. In these studies, the author systematized the theoretical basis of IRT, and provided guidance to the practice of analyzing multiple-choice items according to IRT based on specialized softwares. A typical event marking a new step in the scientific field of educational measurement and evaluation in Vietnam is the introduction of VITESTA software with the function of analyzing and evaluating multiple-choice items (Lam et al, 2007). Using this software has helped users analyze multiple-choice item tests according to the IRT with one-parameter, two-parameter, and three-parameter models. Besides, the analysis results from VITESTA software also help to introduce the parameters of difficulty and discrimination of the items based on CTT. Up to now, this is the only Vietnamese software capable of performing these specialized analytical functions.

Directly related to the field of quality evaluation of multiple-choice items, in Vietnam there have been some authors at universities who take interest in such research topic. Most of the researches apply CTT or IRT to analyzing and evaluating multiple-choice items with different methods. Specifically, using the PROX method to calibrate the difficulty of multiple-choice

items and examinees' ability (Nguyen & Nguyen, 2006), the application of the Gibbs sampling method to estimating the difficulty of items in the Rasch model (Le et al, 2017), the application of IATA software to analyze, evaluate and improve the quality of multiple-choice items Bui & Bui, 2018; Pham & Bui, 2019; Nguyen & Nguyen, 2020, the application of software R (*ltm* package) with three-parameter model to measure the difficulty and discrimination of items on multiple-choice item test, and at the same time survey the effect of the predictive level of examinees when answering items with the examinee ability's measurement and evaluation (Doan et al, 2016). Besides, some authors have used Quest/Conquest software to analyze and evaluate multiple-choice items based on IRT with one-parameter and two-parameter models Nguyen, 2008; Bui, 2017; Nguyen & Nguyen, 2020. In addition, in recent times, there have been some studies to analyze and evaluate multiple-choice items through the combination of S-P chart, gray relationship analysis, and ROC curve (Nguyen, 2015), applying GSP chart and ROC method in combination with assessment based on IRT (Nguyen, 2017).

Most studies that apply IRT to the analysis of multiple-choice test are only use one of three models (one-parameter, two-parameter, three-parameter). In this article, we will use a combination of two-parameter and three-parameter models at the same time to evaluate multiple-choice test.

3. Theoretical basic and research methodology

3.1. Item Response Theory

IRT is a theory of measurement science in education, launched in the 1970s and has been developed strongly to date. This theory builds mathematical models to process data based on the study of every interaction pair between "examinee-item" when implementing an objective test (Lam, 2011, p.82). How each examinee respond to an item will depend on their potentialities and the characteristics of the item. IRT consists of three common mathematical models corresponding to the number of parameters used in the model.

The simplest model of IRT is one-parameter, also known as the Rasch model, which is based on the Rasch view as follows:

"A person having a greater ability than another person should have the greater probability of solving any item of the type in question and similarly, one item being more difficult than another means that for any person the probability of solving the second item correctly is the greater one" (Rasch, 1960, p.117)

In this model, to consider the relationship between the examinee-the item, Rasch selects the ability parameter for each examinee and the difficulty parameter for each item. The mathematical equation for the Rasch model is given below:

$$P(\theta) = \frac{1}{1+e^{-(\theta-b)}} \qquad (1)$$

where: e is the constant 2.718, b is the difficulty parameter of item, θ is the ability level, and $P(\theta)$ is the probability of correct response of examinee who has the ability level of θ.

The difficulty parameter, denoted by b, is defined as the point on the ability scale at which the probability of correct response to the item is 0.5. This is a characteristic parameter for the examinee's ability to answer items correctly, the higher the difficulty of an item, the lower the probability of answering the item correctly. The theoretical range of the values of the b parameter is $-\infty<b<+\infty$. However, typical values of the parameter b is $-3\leq b\leq 3$ (Baker, 2001, p.168).

The curve that represents the characteristic function of the item is called the *item characteristic curve*. For the one-parameter model, the item characteristic curve looks like Figure 1.

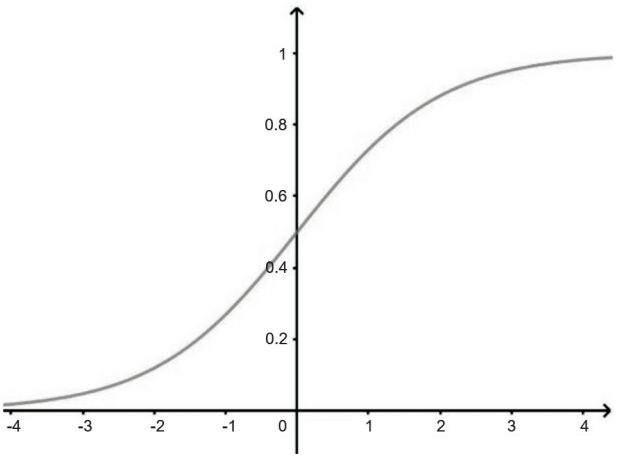

Figure 1. The item characteristic curve following one-parameter model with $b = 0$

The curve represents an examinee's probability of giving the correct answer to that item, based on the examinee's ability. The higher the ability of the examinee is, the greater the probability of correctly answering the item of the examinee will become, and this probability progresses to value 1 when the examinee's ability reaches infinitely positive.

Based on the one-parameter model, with each multiple-choice item in the test, in addition to the parameter, Birnbaum (1968) proposed extending more one parameter, the discrimination parameter, to show the ability to examinee's classification (Doan et al, 2016). This model is called two-parameter model. The equation for the two-parameter model is seen below:

$$P(\theta) = \frac{1}{1+e^{-a(\theta-b)}} \qquad (2)$$

where: e is the constant 2.718, b is the difficulty parameter, a is the discrimination parameter and θ is the ability level.

The discrimination parameter of the item shows the ability to classify examinees for taking the test. The higher the discrimination of the item is, the greater the difference in the probability of getting the correct answer between the high and low examinees will become. The theoretical value range of the parameter a is $-\infty<a<+\infty$. However, typical values are $0.5\leq a\leq 2.0$ (Baker, 2001, p.168). Items that have the parameter a too large or too small are often not significant in measuring the examinees' ability.

For the two-parameter model, the item characteristic curve looks like Figure 2.

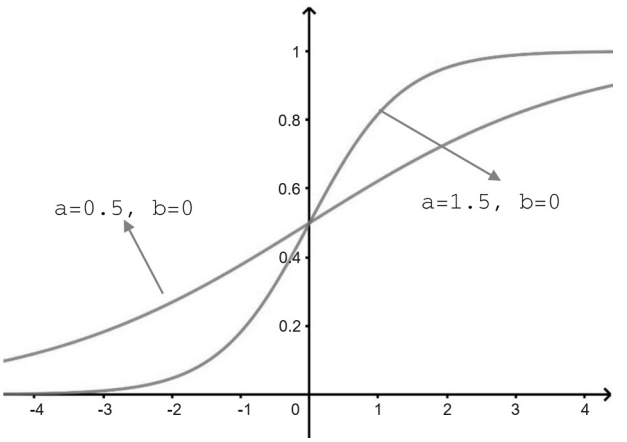

Figure 2. The item characteristic curve following one-parameter model

Compared with the one-parameter model, when the same parameter b value, the item characteristic curve in the two-parameter model has a greater slope when the parameter value a > 1, and has a smaller slope when parameter a value < 1. The steeper the characteristic curve of the item is, the higher the discrimination of the item will become.

In fact, in the process of doing tests, some examinees may select items correctly only by personal sheer guessing. Thus, Birnbaum (1968) proposed adding a guessing parameter $c\in(0,1)$ to the two-parameter model to form three-parameter model (Doan et al, 2016). The equation for the three-parameter model is given below:

$$P(\theta) = \frac{1 - c}{1+e^{-a(\theta-b)}} \qquad (3)$$

where: e is the constant 2.718, b is the difficulty parameter, a is the discrimination parameter, c is the guessing parameter and θ is the ability level.

The parameter c is the probability of getting the item correct of examinee by guessing. Thus,

theoretical range of the *c* parameter is (0,1), but in practice, the parameter *c* that should be used is $0 \le c \le 0.35$ (Baker, 2001, p.168).

The item characteristic curve according to the three-parameter model looks like Figure 3

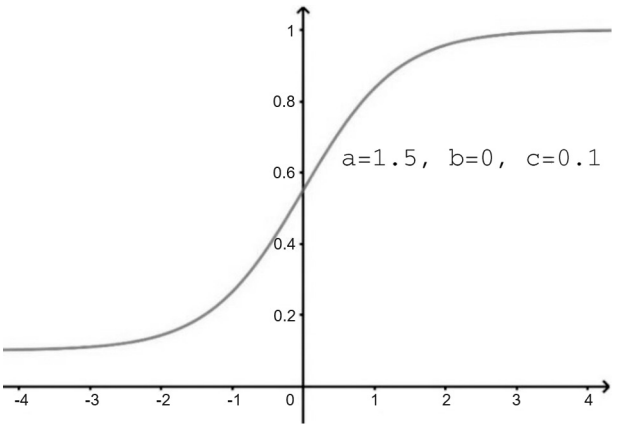

Figure 3. The item characteristic curve following three-parameter model

The item characteristic curve according to the three-parameter model shows when the examinee's ability parameter is very low and progresses to extremely negative, the probability of correctly answering this item does not progress to 0, this probability value approaches the guessing parameter value of that item.

Thus, the higher the guessing value of the item is, the greater the correct answering probability of the examinee to that item will become. This factor will reduce accuracy when assessing the examinee's ability because the correct answer to items with high guess value is due to random factors, not by being influenced by the examinee's ability.

3.2. Introducing R software and using *ltm* package

R software is one of the most popular statistical softwares in the world. One of the functions of this software is to analyze multiple-choice item according to IRT. With the use of package *ltm*, R software will analyze multiple-choice items according to one-parameter, two-parameter, and three-parameter models based on the test taker's response data for the test (Rizopoulos, 2006). To use the *ltm* package, the user needs to install this package and some support packages such as *mirt, mvtnorm, msm* with the following command lines:

```
install.packages("ltm")
install.packages("mirt")
install.packages("mvtnorm")
install.packages("msm")
library(ltm)
```

In addition, the user should prepare the examinee's test data, in which correct answers will be encoded as 1, and incorrect answers will be encoded as 0 (Table 1). R software can read some of data formats, one of which is *.csv format through read.csv() command with the following structure:

```
Data=read.csv("D:/Data.csv", header=T)
```

The data analysis of each model depends on the command lines used, as follows:

For the one-parameter model (Rasch model):

```
Model1PL <- rasch(Data,IRT.param=T)
coef(Model1PL, prob = T, order = T)
```

For the two-parameter model:

```
Model2PL=ltm(Data~z1,IRT.param=T)
Summary(Model2PL)
coef(Model2PL)
```

For the three-parameter model:

```
Model3PL = tpm(Data, type="latent.trait", IRT.param=T)
Summary(Model3PL)
coef(Model3PL)
```

Where, the command line coef () will help display the parameters of multiple-choice items according to the corresponding model. Specifically, the difficulty parameter for the Rasch model; difficulty and discrimination parameters for two-parameter model; difficulty, discrimination and guessing parameters for three-parameter model.

3.3. Research data

Data used in this study are the responses by 590 students on English 1 test in Dong Thap University in 2018. The test consists of 50 multiple-choice items; each item has 04 answer options, including 01 correct option and 03 interfering options. A part of the students' answers to the test is shown in Table 1.

Table 1. Extract a part of the data

ID	Item 1	Item 2	Item 3	Item 4	…	Item 47	Item 48	Item 49	Item 50
1	1	0	1	1	...	1	1	0	1
2	0	1	1	0	...	1	1	0	0
3	1	0	1	0	...	1	0	0	1
4	0	0	1	0	...	1	1	0	0
5	0	0	1	0	...	1	1	0	1
6	0	0	0	0	...	1	1	0	0
7	0	0	1	0	...	1	1	0	0
8	0	0	0	1	...	1	1	1	1
9	0	0	0	1	...	1	0	1	0
10	1	1	1	0	...	1	1	0	0
11	0	1	1	1	...	1	1	0	1
...	...	...	...	...	...	...	...	...	...
581	0	0	0	1	...	1	1	1	0
582	0	1	1	0	...	1	1	1	0
583	1	1	1	1	...	1	1	1	1
584	0	1	1	1	...	1	1	0	1
585	0	0	1	0	...	1	1	1	0
586	0	1	1	1	...	1	1	1	0
587	1	0	1	0	...	1	1	0	0
588	0	1	1	0	...	0	1	0	0
589	0	0	1	1	...	1	1	1	1
590	0	1	1	1	...	1	1	1	1

The results of each student's response to each item are coded into the values 0 and 1. Of which, value 1 represents the correct response, and value 0 represents the incorrect response for each item. This data format is also required for statistical analysis with R software.

3.4. The reliability of the data

Before using R software, we have conducted to assess the reliability of the data through Cronbach Alpha coefficients. The result of calculating Cronbach Alpha value is 0.796, which shows that the data have a high level of reliability, suitable for conducting further analysis and evaluation.

4. Research results and discussions

To perform analysis, evaluation, and selection of multiple-choice items, we will use the calculation results of the difficulty parameters *a*, discrimination *b*, and the guessing parameter *c* of the items according to two-parameter and three-parameter models based on IRT from R software. Specifically, the two-parameter model is used first to evaluate the items based on the difficulty and discrimination parameters. Next, the three-parameter model is used to evaluate items with guessing parameter c next to difficulty and discrimination parameters. In reality, doing the test, many examinees cannot determine the correct

answer to some items but they can give correct responses to these items due to random factors, not entirely due to the examinee's true ability.

Analysis results of multiple-choice items according to the two-parameter model using R software (*ltm* package) are shown in Table 2.

Table 2. Data analysis results according to the two-parameter model

Item	Parameters		Item	Parameters		Item	Parameters	
	a	*b*		*a*	*b*		*a*	*b*
1	0.52	0.53	18	0.33	0.08	35	-0.03	-32.84
2	0.66	-0.19	19	0.93	0.46	36	0.83	-0.46
3	1.28	-1.93	20	0.34	0.59	37	0.55	2.27
4	0.81	-0.62	21	1.40	-0.26	38	0.45	2.19
5	0.47	0.05	22	0.79	0.19	39	0.23	-0.09
6	0.82	0.40	23	0.69	1.22	40	0.85	0.44
7	1.00	-0.33	24	0.46	0.30	41	0.29	1.72
8	1.74	-0.97	25	0.50	2.55	42	0.59	0.94
9	1.32	0.38	26	0.77	-0.63	43	-0.03	-51.50
10	-0.46	-1.95	27	1.06	-1.06	44	0.61	0.77
11	0.98	-0.76	28	0.17	5.15	45	0.40	2.97
12	1.17	0.58	29	1.31	-0.26	46	1.04	0.87
13	0.63	-1.28	30	0.17	0.62	47	0.94	-2.22
14	0.07	10.05	31	0.94	1.31	48	0.31	-7.56
15	0.87	-0.40	32	1.08	-1.13	49	0.24	3.96
16	0.82	-1.91	33	1.24	-0.97	50	0.90	0.88
17	0.35	3.41	34	1.05	-0.50			

The calculation of the value of the difficulty and discrimination parameters of multiple-choice items according to IRT can reach values from $-\infty$ to $+\infty$. However, the items that consist of very low or very high parameter values often have no meaning to provide information to measure and evaluate the examinee's performance. Therefore, in order for evaluating and selecting appropriate items to be grounded, we use commonly-used ranges of values for the parameters of multiple-choice items as proposed by Baker (2001). Specifically, the values of the difficulty parameter *b* and the discrimination parameter *a* often meet $-3.0 \le b \le 3.0$ and $0.0 \le a < 2.0$. Thus, an item has good quality when both the difficulty parameter and discrimination parameter are in the proposed range as above. Conversely, an item is not qualified when it contains at least one of the two parameters that their values are outside the recommended range. With the above evaluation, this test contains some suboptimal items that need to be reviewed before they are put into item bank. These items are shown in Table 3

Table 3. The items are not good when considering the two-parameter model

Item	Parameters		Item	Parameters		Item	Parameters	
	a	*b*		*a*	*b*		*a*	*b*
5	0.47	0.05	24	0.46	0.30	41	0.29	1.72
10	-0.46	-1.95	28	0.17	5.15	43	-0.03	-51.50
14	0.07	10.05	30	0.17	0.62	45	0.40	2.97
17	0.35	3.41	35	-0.03	-32.84	48	0.31	-7.56
18	0.33	0.08	38	0.45	2.19	49	0.24	3.96
20	0.34	0.59	39	0.23	-0.09			

The analysis of multiple-choice items according to the two-parameter model shows that there are 17 unsatisfactory items in this test that need to be considered. Specifically, 10 items (5, 10, 18, 20, 24, 30, 38, 39, 41, 45) contain one unsatisfactory parameter and 7 items (14, 17, 28, 35, 43, 48, 49) contain both unsatisfactory parameters. In fact, the results of answering each multiple-choice item of examinees are also influenced by guessing factor because examinees choose the answers randomly. Therefore, to ensure the evaluation of items to be more comprehensive, we continue to analyze each item according to the three-parameter model of IRT.

Analysis results of items according to the three-parameter model using R software (*ltm* package) are shown in Table 4.

Table 4. Data analysis results according to the three-parameter model

Item	Parameter			Item	Parameter		
	a	***b***	***c***		***a***	***b***	***c***
1	1.54	1.53	0.33	26	1.65	0.61	0.39
2	0.71	-0.12	0.00	27	1.56	-0.12	0.37
3	1.22	-2.00	0.00	28	0.52	4.39	0.21
4	0.79	-0.54	0.02	29	1.97	0.26	0.21
5	0.73	1.21	0.26	30	0.29	3.76	0.29
6	1.28	0.97	0.19	31	2.62	1.40	0.15
7	4.62	0.80	0.42	32	3.51	0.32	0.54
8	2.13	-0.50	0.27	33	1.83	-0.14	0.36
9	2.62	0.76	0.17	34	2.63	0.54	0.39
10	-0.41	-2.17	0.00	35	-1.08	-3.03	0.23
11	1.31	-0.02	0.28	36	3.82	0.89	0.46
12	2.72	0.94	0.18	37	1.30	2.08	0.14
13	0.82	-0.17	0.31	38	1.28	2.22	0.20
14	1.55	2.71	0.31	39	2.17	1.95	0.47
15	0.94	-0.13	0.08	40	1.56	0.99	0.21
16	5.09	0.64	0.72	41	2.89	2.03	0.35
17	1.46	2.45	0.19	42	0.66	1.10	0.04
18	0.36	0.14	0.01	43	-0.04	48.58	0.04
19	1.00	0.50	0.00	44	1.18	1.30	0.21
20	0.37	0.70	0.02	45	1.94	2.17	0.19
21	1.35	-0.19	0.00	46	1.52	1.08	0.10
22	2.60	1.00	0.32	47	2.00	0.12	0.73
23	1.40	1.52	0.17	48	0.26	-8.45	0.07
24	0.57	0.78	0.11	49	0.60	3.67	0.19
25	1.18	2.23	0.14	50	2.16	1.20	0.18

Analyzing multiple-choice items according to the three-parameter model, users can evaluate guessing parameter c ($0 \le c \le 1$) of each item. According to Baker (2001), the value of guessing parameter c is not greater than 0.35. Thus, when evaluating the item according to the three-parameter model, an item is good when all three parameters a, b, c are in the ranges $-3 \le b \le 3$, $0.5 \le a < 2$, and $0 \le c \le 0.35$. Besides, an item that is not good needs to be further considered for correction when it contains at least one parameter outside of the above ranges. By identifying the way of evaluating the item according to the three-parameter model, this test of 26 bad items needs further consideration. These items are shown in Table 5.

Table 5. The items are not good when considering the three-parameter model

Item	Parameter			Item	Parameter		
	a	*b*	*c*		*a*	*b*	*c*
7	4.62	0.80	0.42	31	2.62	1.40	0.15
8	2.13	-0.50	0.27	32	3.51	0.32	0.54
9	2.62	0.76	0.17	33	1.83	-0.14	0.36
10	-0.41	-2.17	0.00	34	2.63	0.54	0.39
12	2.72	0.94	0.18	35	-1.08	-3.03	0.23
16	5.09	0.64	0.72	36	3.82	0.89	0.46
18	0.36	0.14	0.01	39	2.17	1.95	0.47
20	0.37	0.70	0.02	41	2.89	2.03	0.35
22	2.60	1.00	0.32	43	-0.04	48.58	0.04
26	1.65	0.61	0.39	47	2.00	0.12	0.73
27	1.56	-0.12	0.37	48	0.26	-8.45	0.07
28	0.52	4.39	0.21	49	0.60	3.67	0.19
30	0.29	3.76	0.29	50	2.16	1.20	0.18

The statistical results in Table 5 show that 10 items contain two unsatisfactory parameters, such as item 7 (a = 4.62, c = 0.42), item 16 (a = 5.09, c = 0.72), item 30 (a = 0.29, b = 3.76), item 32 (a = 3.51, c = 0.54), item 34 (a = 2.63, c = 0.39), item 35 (a =-1.08, b = -3.03), item 36 (a = 3.82, c = 0.46), item 39 (a = 2.17, c = 0.47), item 43 (a = -0.04 , b = 48.58), and item 48 (a = 0.26, b = -8.45). Besides, the remaining 16 items all contain one unsatisfactory parameter.

Thus, the results of analyzing the items according to the two-parameter model and the three-parameter model resulted in different evaluation results. Specifically, using the three-parameter model was shown 26 bad items compared to 17 bad items that considered under the two-parameter model because using the three-parameter model with a guessing parameter has influenced the difficulty and discrimination parameters of each item. The evaluation results for each model are shown in Table 6.

Table 6. Evaluation results of the test according to two-parameter model and three-parameter model

Items	Evaluation result		Items	Evaluation result	
	Two-parameter model	Three-parameter model		Two-parameter model	Three-parameter model
1	Good	Not Good	26	Good	Not Good
2	Good	Good	27	Good	Not Good
3	Good	Good	28	Not Good	Not Good
4	Good	Good	29	Good	Good
5	Not Good	Good	30	Not Good	Not Good
6	Good	Good	31	Good	Not Good
7	Good	Not Good	32	Good	N Not Good
8	Good	Not Good	33	Good	Not Good
9	Good	Not Good	34	Good	Not Good
10	Not Good	Not Good	35	Not Good	Not Good
11	Good	Good	36	Good	Not Good
12	Good	Not Good	37	Good	Good

Items	Evaluation result		Items	Evaluation result	
	Two-parameter model	Three-parameter model		Two-parameter model	Three-parameter model
13	Good	Good	38	Not Good	Good
14	Not Good	Good	39	Not Good	Not Good
15	Good	Good	40	Good	Good
16	Good	Not Good	41	Not Good	Not Good
17	Not Good	Good	42	Good	Good
18	Not Good	Not Good	43	Not Good	Not Good
19	Good	Good	44	Good	Good
20	Not Good	Not Good	45	Not Good	Good
21	Good	Good	46	Good	Good
22	Good	Not Good	47	Good	Not Good
23	Good	Good	48	Not Good	Not Good
24	Not Good	Good	49	Not Good	Not Good
25	Good	Good	50	Good	Not Good

The statistical results in Table 6 show that some items are *good* when they are evaluated according to the two-parameter model, but they are *not good* when considered according to the three-parameter model, and vice versa. Besides, using the ltm package, the software R also shows the compatibility between the two-parameter and three-parameter models for the analyzed data. These statistical results are shown in Table 7.

Table 7. Suitability between models with data

Likelihood Ratio Table						
	AIC	BIC	log.Lik	LRT	df	p.value
Model2PL	34782.95	35220.97	-17291.48			
Model3PL	34716.60	35373.61	-17208.30	166.36	50	<0.001

Based on the model selection theory, the better model is the one with smaller AIC, BIC, and log.Lik indicators simultaneously (Rizopoulos, 2006). However, the statistical results in Table 7 for AIC, BIC, and log.Lik values in the two-parameter model is not simultaneously smaller or larger than these values in the three-parameter model. Therefore, it is not possible to evaluate a better fit for the data between the two-parameter model and the three-parameter model. It shows that each model has its advantages in evaluating multiple-choice items. Thus, to evaluate the items more comprehensively, helping to choose the optimal items, we propose the selection of well-evaluated items according to both two-parameter mode and three-parameter model. By choosing the items as above, 17 satisfactory items in the test, such as item 2, item 3, item 4, item 6, item 11, item 13, item 15, item 19, item 21, item 23, item 25, item 29, item 37, item 40, item 42, item 44, and item 46 could meet the testing requirements, so they can be put into item banks. In addition, the remaining items need to be reviewed before being put to use.

The researcher have continued to survey 15 English majored lecturers (in Dong Thap University) for their comments on the test as a reference for comparing the results of analysis, which is based on IRT, on the students' responses to the test of English one.

The evaluation of each item on the test is done by the lecturers' remarks on the item's level of difficulty (very easy, easy, medium, difficult and very difficult) and the degree of discrimination (very poor, poor, average, good and very good). Accordingly, the lecturers will make the final conclusion of *not good* - need further amendment or *good* - can be put into use for each item. The results of a detailed evaluation by the lecturers on the test items are shown in Table 8.

Table 8. Results of lecturers' comments on the test items

Number of valuation		Number of item	Items
Not good	Good		
0	15	14	9, 15, 18, 19, 20, 24, 27, 29, 30, 31, 32, 34, 47, 49
1	14	14	2, 7, 11, 14, 16, 17, 21, 33, 38, 40, 44, 46, 48, 50
2	13	10	4, 5, 10, 12, 25, 26, 35, 37, 39, 42
3	12	5	6, 23, 41, 43, 45
4	11	3	1, 8, 22
5	10	1	36
6	9	2	13, 28
7	8	1	3

The evaluation results show that among the test items, 14 items received 15/15 good reviews from the lecturers, including items 9, 15, 18, 19, 20, 24, 27, 29, 30, 31, 32, 34, 47, 49. The remaining 36 items all received *not good* feedback from the lecturers. Specifically, the number of not good responses from the lecturers to the above items accounted for from 1 to 7 times out of 15 lecturers, from 6.3% to 46.7% respectively. In addition, the statistical results in Table 8 showed that among 33 *not good* items, when analyzed and evaluated by the IRT models (Table 6), 22 items were rated as *not good* with the number of responses from 1 to 6 times out of 15 lecturers, including the items 1, 5, 7, 8, 10, 12, 14, 16, 17, 22, 26, 28, 33, 35, 36, 38, 39, 41, 43, 45, 48, 50. However, among the items rated *not good* based on the IRT (Table 6), 11 items that did not receive a n*ot good* rating from the lecturers, namely items 9, 18, 20, 24, 27, 30, 31, 32, 34, 47, 49. Furthermore, among 17 items rated as good by the IRT, 14 items received a *not good* rating, specifically the item 1, 2, 3, 6 and 7 votes respectively by 15 lecturers. Thus, only 3 items rated *good* by IRT did not receive *not good* rating from the lecturers (items 15, 19 and 29).

This shows that it is possible that the *not good* items could be introduced into the test without being detected if the evaluation of the test items is merely based on the subjective evaluation by professional lecturers, not on the process of analyzing and evaluating test takers' results based on IRT statistics models. At that time, the assessment results of the students' ability will not be totally objective and accurate. Therefore, in order for the test items to be able to accurately evaluated and the *good* items to be chosen, it is necessary to conduct the process of analyzing and evaluating test items based on the statistical models of the IRT through specialized software in addition to the reference of the professional lecturers' comments. This combination will help to comprehensively evaluate the items before introducing them into the item banks and using in the tests. This will enable the assessment results to be more objective, and the assessment of the students' capacity to be more accurate.

5. Conclusion

Analyzing and evaluating multiple-choice items based on difficulty, discrimination, and guessing parameters according to the two-parameter and three-parameter models of IRT have shown *good* items and *not good* items. Of which, the *good* items can put into the item bank to use in the assessment of learning results, the *not good* items need to be further considered before being put into use. This shows that evaluating each item by experts and especially quantitative analysis

based on specialized software will be of essential necessity. The analysis will help the test editor determine the parameters of each multiple-choice item. On that basis, the teachers can actively choose appropriate items to put in the tests and this helps assess learners' ability accurately. In particular, teachers can design equivalent tests to use in different exams based on the estimated parameters of each item. This helps teachers attain fair and objective assessment, contributing to achieving the goals of teaching activities./.

Acknowledgement: This research is supported by science and technology project, Dong Thap University. Code: SPD2020.01.36

References

Baker, F. B. (2001). T*he Basics of Item Response Theory*, ERIC Clearinghouse on Assessment and Evaluation, College Park, MD.

Bui, N. Q. (2017). Evaluation of the quality of multiple choice test bank for the module of Introduction to Anthropology by using the RASCH model and QUEST software. *Science of Technology Development*, *20 (X3)*, 42-54.

Bui, A. K., & Bui, N. P. (2018). Using IATA to analyze, evaluate and improve the quality of the multiple-choice items in chapter power functions, exponential functions and logarithmic functions. *Can Tho University Journal of Science*, *54(9C)*, 81-93.

Doan ,H. C., Le, A. V., & Pham, H. U. (2016). Applying three-parameter logistic model in validating the level of difficulty, discrimination and guessing of items in a multiple choice test. *Ho Chi Minh city University of Education Journal of Science*, *7(85)*, 174-184.

Duong, T. T. (2005). *Test and measure academic achievement*. Hanoi: Social Sciences Publishing House.

Lam, Q. T., Lam, N. M., Le, M. T., & Vu, D. B. (2007). VITESTA software and analysis of test data. *Vietnam Journal of Education*, *176*, 10-12.

Lam, Q. T. (2011). *Measurement in Education - Theory and Application*. Hanoi: Vietnam National University Publishing House.

Le, A. V., Pham, H. U., Doan, H. C., & Le, T. H. (2017). Using Gibbs Sampler to evaluate item difficulty in Rasch model. *Ho Chi Minh city University of Education Journal of Science*, *14(4)*, 119-130.

Rizopoulos, D. (2006). An R package for latent variable modeling and item response theory analysis. *Journal of Statistical Software*, *17(5)* 1-25.

Nguyen, B. H. T. (2008). Using Quest software to analyze objective test questions. *Journal of Science and Technology*, *Da Nang University*, 2, 119-126.

Nguyen, P. H. (2017). Using GSP chart and ROC method to analyze and select mutiple choice items. *Dong Thap University Journal of Science*, *24 (2)*, 11-17.

Nguyen, P. H., & Du, T. N. (2015). The analysis and selection of objective test items based on S-P chart, Grey Relational Analysis, and ROC curve. *Ho Chi Minh City University of Education Journal of Science*, *6(72)*, 163-173.

Nguyen, T. H. M., & Nguyen, D. T. (2006). Measurement Assessment in the objective test: Question difficulty and Examinees' ability. *Vietnam National University Journal of Science, 4*, 34-47.

Nguyen, V. C., & Nguyen, Q. T. (2020). Applying ConQuest software with the two-parameter IRT model to evaluate the quality of multiple-choice test. *HNUE Journal of Science*, *65(7)*, 230 - 242.

Pham, T. M., & Bui, D. N. (2019). The IATA software for analyzing, evaluation of multiple-choice questions at Ha Noi Metropolitan University. *Scientific Journal of Ha Noi Metropolitan University*, *20*, 97-108.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.