ANALYSIS AND EVALUATION OF MULTIPLE-CHOICE TEST ITEMS AND TEST DESIGN: A STUDY ON APPLICATION OF ITEM RESPONSE THEORY

Nguyen Van Canh[*], Pham Van Tac
Dong Thap University

ARTICLE INFO

Received:	**12/8/2021**
Revised:	**07/9/2021**
Published:	**07/9/2021**

KEYWORDS

Item Response Theory
Multiple – choice item
Information Curve
R software (package ltm)
GeoGebra software

ABSTRACT

This article presents the findings on applying the item response theory for a 2-parameter model in analyzing and evaluating question items and designing multiple-choice tests. Based on the results of data analysis by R (package ltm) of non-majored students' performance in English 1 exam papers used in Dong Thap University from 2017 to 2021, the study identified the satisfactory items which could meet the exam requirements and unsatisfactory question ones for further adjustment or improvement. Among the unsatisfactory items, some abnormal characteristics, seriously violating the tests' limitations in terms of difficulty and discriminate level, must have been definitely removed from the test papers. In addition, the study findings also show that the test items had a relatively low level of measuring students' competence (below 0.0 according to the competence scale). Finally, the study introduced the way of applying the information curve tool to design test items to help accurately measure the students' competence based on the characteristic parameters of items.

PHÂN TÍCH, ĐÁNH GIÁ CÂU HỎI TRẮC NGHIỆM KHÁCH QUAN VÀ XÂY DỰNG ĐỀ THI: MỘT NGHIÊN CỨU ỨNG DỤNG LÝ THUYẾT ỨNG ĐÁP CÂU HỎI

Nguyễn Văn Cảnh[*], Phạm Văn Tặc
Trường Đại học Đồng Tháp

THÔNG TIN BÀI BÁO

Ngày nhận bài:	**12/8/2021**
Ngày hoàn thiện:	**07/9/2021**
Ngày đăng:	**07/9/2021**

TỪ KHÓA

Lý thuyết ứng đáp câu hỏi
Câu hỏi trắc nghiệm khách quan
Đường cong thông tin
Phần mềm R (gói ltm)
Phần mềm GeoGebra

TÓM TẮT

Bài viết trình bày kết quả ứng dụng lý thuyết ứng đáp câu hỏi với mô hình 2 tham số vào việc phân tích, đánh giá câu hỏi và xây dựng đề thi trắc nghiệm khách quan. Dựa trên việc phân tích dữ liệu kết quả thi của sinh viên (không thuộc chuyên ngành Tiếng Anh) đối với các đề thi Tiếng Anh 1 được sử dụng tại Trường Đại học Đồng Tháp từ năm 2017 đến 2021 bằng phần mềm R (gói ltm), nghiên cứu đã chỉ ra những câu hỏi đạt yêu cầu, đủ điều kiện để sử dụng trong các đề thi và những câu hỏi chưa đạt yêu cầu, cần phải được xem xét lại để điều chỉnh, cải tiến. Trong đó, một số câu hỏi được sử dụng trong các đề thi có dấu hiệu bất thường, vi phạm nghiệm trọng về giới hạn giá trị các tham số độ khó, độ phân biệt cần phải được loại bỏ ra khỏi đề thi. Ngoài ra, kết quả nghiên cứu còn cho thấy các đề thi trên đều có ý nghĩa đo lường mức năng lực khá thấp (dưới 0.0 theo thang đo năng lực). Bên cạnh đó, nghiên cứu đã giới thiệu cách vận dụng công cụ đường cong thông tin vào việc xây dựng các đề thi giúp đo lường chính xác năng lực của người học dựa trên các tham số đặc trưng của các câu hỏi.

DOI: https://doi.org/10.34238/tnu-jst.4885

[*] Corresponding author. *Email: nvcanh@dthu.edu.vn*

1. Introduction

The scientific field of measurement and assessment in education began to take its form and thrived around the 1970s by the birth and improvement of the Classical Test Theory (CTT). This is one of the theories making many important contributions to the work of measurement and evaluation activities in education, especially in the analysis and assessment of exam questions. However, this theory has some limitations, one of which cannot separate the characteristics of independent test takers from the characteristics of the multiple-choice items, with the former being able to be explained in relation to the latter's features [1]. To overcome the above limitations of CTT, Rasch suggested that the analysis and evaluation of multiple choice questions was only valid when it was based on each individual test taker, in which the test taker's characteristics were separated from the questions [2]. This Rasch's viewpoint marked a transition from the CTT model to the Item Response Theory (IRT) model, a mathematical model that describes the probability of students' answering questions correctly in the corresponding level between the test takers' competence and the difficulty of the questions. The mathematical formula of this model is shown in the following form:

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

with *e* being a constant 2.718, *b* being the difficulty parameter of the test item, θ the parameter of the candidate's ability and $P(\theta)$ the probability of answering the question correctly by the test takers with their competence level of θ. In the Rasch model, if the test takers' competence is equal to the difficulty of a question, their probability of answering the question correctly is 50%. On the basis of Rasch model, Birnbaum proposed to extend the discrimination parameter *a* of the item to show the possibility of candidates [3] as in the following formula:

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

During a multiple-choice test, some test takers are able to answer the items correctly based on random predictions. Therefore, Birnbaum [3] proposed adding the prediction parameter *c* of the question to the 2-parameter model to form a 3-parameter model as in the following formula:

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

With the appearance of the guessing parameter *c* $\left(c \in (0,1)\right)$, and their very low ability, the test takers' probability of correctly answering the questions does not move towards the value 0 but towards the value of the guessing parameter *c* of that test items.

On the basis of IRT, many studies have been carried out to achieve different goals. In Baker's study, an item in multiple-choice tests is divided into five levels of difficulty: very easy, easy, medium, difficult, very difficult; at the same time 5 levels of quality discrimination: very poor, poor, average, good, very good. In addition, the author proposed a value limit for the parameters of the items used in the exam. Specifically, the difficulty of the items should be from -3.0 to 3.0; The discrimination should be from 0.5 to less than 2.0 and the prediction value should be from 0 to less than 0.35 [4]. Bortolotti and his research group members presents in their study the basic and fundamental concepts of IRT and a practical example about proposing the construction of scales to illustrate the feasibility, advantages and validity of IRT through a known measurement. The obtained results from the practical application of IRT confirm its effectiveness in the evaluation of intangible traits [5]. Furthermore, another study by Rakkapao revealed that IRT analysis is useful in assessing the test since its item parameters are independent of the students' competence parameters. Moreover, the item response curves analysis can be used to assess the effectiveness of the test's distractors. Test developers can apply these methods to diagnose and evaluate the features of items at various ability levels of test takers [6].

In term of level diagnostic statistics and model-data fit with 1 and 2-parameter models using IRTPRO V3.0 and BILOG-MG Version 3.0, Essen recommended that the use of more than one IRT softwares offers more useful information for the choice of a model that fits the data [7]. Similarly, Foster identified and coded 63 articles that used IRT on empirical data published in industrial-organizational and organizational behavior journals since 2000. Results show that typical usage for IRT conforms to best practices in several ways; however, in other ways, such as testing for and reporting the appropriateness of the given models, there have remained significant limitations for further improvement [8]. Setiawati described the items' parameters analysis result in relation to measurement. The items' parameters analyzed in this instrument met the appropriateness of models, items' difficulty, items' discriminimant, items' prediction parameters, items' information curves, and test information function [9]. In addition, Mu'iz had a study to clarify the characteristics of a multiple-choice test in terms of its validity, reliability, discriminant, difficulty and prediction parameter based on applying IRT into measuring students' critical thinking level and masterfulness of concepts [10].

In Vietnam, the scientific field of measurement and evaluation in education was formed quite late and has developed more slowly than those in many countries in the world. A typical event marking a new step of this science in Vietnam is the introduction of VITESTA software with the function of analyzing and evaluating multiple-choice tests based on IRT and CTT [11]. In addition, several studies related to the evaluation of objective multiple-choice tests have been carried out by different approaches and methods. Specifically, there was the use of PROX method, which is a measurement method based on Rasch model to size the difficulty of multiple-choice items and to evaluate examinees' ability [12] and the application of Gibbs sampling method to estimating the difficulty of the test items by Rasch model [13]. Some related studies were the application of IATA software to analyze, evaluate and improve the quality of multiple-choice tests [14], [15] and the application of R software (package ltm) with 3-parameters model to measure the difficulty, discriminant level of the test items in multiple-choice tests, and at the same time to investigate the influence of the students' prediction level on their answering the tests in assessing students' competence [16]. Other studies were done by using Quest/Conquest software to analyze and evaluate multiple-choice questions based on IRT [17] - [19]. Finally, there were a number of studies on analyzing and evaluating multiple-choice items through the combined use of SP (Student-Problem) chart, analysis of gray relationship and ROC (Receiver Operating Characteristics) curves [20], the application of GSP (Grey Student-Problem) chart and ROC method combined with assessment based on IRT [21].

The work of analyzing and evaluating multiple-choice test items in the above studies have shown certain advantages in making recommendations to selecting satisfactory items, as well as pointing out unsatisfactory ones in exam papers. However, very few studies have referred to the application of IRT to writing multiple-choice tests capable of accurately measuring students' competence. This study was conducted for two main objectives: (1) Analyzing and evaluating multiple-choice tests by using IRT (through data analysis results from R software), thereby indicating the quality of of the used test questions; (2) proposing a way to determine the information curve of the multiple-choice tests by using IRT and using the obtained results to write satisfactory multiple-choice items so as to introduce them into the exam papers, thus enabling the users to accurately assess the students' competence and thereby achieving the ultimate goal of the assessment.

2. Research methods

2.1. Research data

The research data used in this study are students' performance results on five different English 1 exam papers organized on Dong Thap University from 2017 to 2021. The above exam papers were designed independently by different lecturers over the school years, so the contents were different, and so was the number of students taking the exam papers. Each exam paper included 50 multiple-choice items, in which each item had 04 answer options including 01 correct option (right answer) and 03 distractors. In addition, the data were evaluated for reliability through Cronbach's Alpha value before being used for statistical analysis for further comments in the study. The results of data reliability analysis are shown in Table 1.

Table 1. *Cronbach's Alpha of the research data*

Exam test	Number of items	Number of candidates	Cronbach's Alpha
2017	50	496	0.872
2018	50	590	0.796
2019	50	876	0.883
2020	50	798	0.807
2021	50	494	0.834

(Source: Analysis results from the authors' data, 2021)

The statistics results in Table 1 show that the number of students taking the English 1 exam paper over the years was quite huge (from 494 to 876 students), and the Cronbach's Alpha reliability value of the test data was from 0.796 to 0.883. This proves that the data used in this study had a high level of reliability and could be used for further analysis.

2.2. Analysis of items in exam papers

The analysis of test data in this study was done through R software. With the ltm package, R software will provide the function of analyzing objective multiple-choice questions based on IRT [22]. In order to use the "ltm" package to analyze the test items and the whole multiple-choice tests, R software requires users to install this package and a number of other support packages such as *mirt, mvtnorm, msm*. In addition, the analysis of the parameters of the multiple-choice item according to the IRT models depends on the command code line used to run the data in the R software. This study applied the 2-parameter model to the analysis of multiple-choice tests with control command lines as follows:

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

Among them, the command coef() helps to display the value of the item's characteristic parameters.

2.3. Drawing the information curve of the multiple-choice tests

2.3.1. The information function of the multiple-choice test

The information function of a multiple-choice test is the total of the information functions of all items of that test [1] and is formed as follow:

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta) \qquad (4)$$

In the above formula, $I(\theta)$ is the function expression of the test and $I_i(\theta)$ is the function expression of the item number *i* used in the multiple-choice test. Birnbaum [3] proposed the function expression of a multiple-choice item as follows.

$$I_i(\theta) = \frac{\left[P'_i(\theta)\right]^2}{P_i(\theta).Q_i(\theta)} \tag{5}$$

Again, in the formulas above $P_i(\theta)$ is the characterized function expression of item number *i* and $Q_i(\theta) = 1 - P_i(\theta)$. For the 2-parameter model, the information function of the item is expressed as follows:

$$I_i(\theta) = \frac{\left[P'_i(\theta)\right]^2}{P_i(\theta).Q_i(\theta)} = (a_i)^2 . \frac{e^{a\left(\theta - b_i\right)}}{\left(1 + e^{a\left(\theta - b_i\right)}\right)^2} \tag{6}$$

In the information function expression, the difficulty parameter value *b* of the item will indicate the level of competence for which the item has the most accurate measurement significance. In addition, the discrimination parameter *a* value of the item will indicate the level of information contribution of the question to the assessment of the students' competence. In a broader meaning, if the item has the greater discrimination parameter *a*, the level of information contribution of that item to the assessment of the students’ competence will be higher in response. Thus, in order to measure the students' competence accurately, test writers should select and introduce the items with difficulty values corresponding to the students’ competence.

2.3.2. Drawing the information curve of the multiple-choice tests

Currently, specialized softwares with the function of analyzing multiple-choice tests based on IRT support the drawing of the information curve for multiple-choice items and the whole test items based on test takers' performance results. However, when we use the characteristic parameters of the items (difficulty, discrimination) in the test to draw the information curve without inputting students’ performance results into the system data, almost all the above softwares do not work. Therefore, the researchers used GeoGebra software to draw the information curve for the multiple-choice items based on the parameters of the question used in the test [23]. The procedure to draw the information curve for a multiple-choice test by GeoGebra was done in three steps and was described in 3.2. In addition, the advantageous point of this software is clearly shown in that users can easily change the items in the test by updating the parameters of those items and the software will quickly display the curve in correspond to the item that has just been updated.

3. Findings and discussions

3.1. Findings on the analysis of the tests by using Item Response Theory

By using R software (package ltm), the researchers were able to show the parameters of each item in respective English 1 exam papers used from 2017 to 2021 at Dong Thap University. On that principle, the evaluation of each multiple-choice item was performed based on its value of the parameters of difficulty, discrimination as proposed by Baker [4]. Specifically, the item was satisfactorily chosen when the difficulty parameter value reached from -3.0 to 3.0, and the discrimination parameter reached a value from 0.5 to less than 2.0. With the above item classification, the number of satisfactory and unsatisfactory items in the English 1 exam papers used over the school years is shown in Table 2.

The results from our statistics in Table 2 show that most of the items used in the English 1 exam papers over the above-mentioned school years had the parameter values of difficulty and discrimination within the acceptable ranges in Baker’s scale ($-3.0 \le b \le 3.0,\ 0.5 \le a < 2.0$) [4].

Table 2. *Description of parameter values for the items in English 1 tests*

Exam test		Difficulty level			Discriminant level		
		b < -3.0	***-3.0 ≤ b ≤ 3.0***	***b > 3.0***	***a < 0.5***	***0.5 ≤ a < 2.0***	***a ≥ 2.0***
2017	Items in total	1	49	0	6	44	0
	Percent %	2.0	98.0	0.0	12.0	88.0	0.0
2018	Items in total	3	43	4	18	32	0
	Percent %	6.0	86.0	8.0	36.0	64.0	0.0
2019	Items in total	0	50	0	2	48	0
	Percent %	0.0	100.0	0.0	4.0	96.0	0.0
2020	Items in total	4	44	2	16	31	3
	Percent %	8.0	88.0	4.0	32.0	62.0	6.0
2021	Items in total	2	45	3	11	38	1
	Percent %	4.0	90.0	6.0	22.0	76.0	2.0

(Source: Analysis results from the authors' data, 2021)

Specifically, in terms of the difficulty parameter, the number of satisfactory items in the exam papers accounted for 43 or more, of which the exam paper in 2019 contained 50 items that met the requirements on the difficulty parameter. In terms of discrimination, the number of satisfactory items in the papers ranged from 31 to 48. Thus, besides the satisfactory items, there were still many unsatisfactory ones in the above exam papers, especially for the discrimination parameter aspect to be taken for consideration. These items are presented in Table 3.

Table 3. *Summary of unsatisfactory item in term of difficulty and discriminant levels*

Exam paper	Unsatisfactory items		
	Item in total	***Difficulty level***	***Discriminant level***
2017	6	43	5, 8, 9, 43, 44, 49
2018	18	14, 17, 28, 35, 43, 48, 49	5, 10, 14, 17, 18, 20, 24, 25, 28, 30, 35, 38, 39, 41, 43, 45, 48, 49
2019	2	None	31, 39
2020	19	4, 8, 32, 38, 40, 44	3, 4, 5, 7, 8, 15, 28, 29, 30, 31, 32, 35, 38, 39, 40, 41, 44, 45, 46
2021	14	24, 30, 39, 41, 46	2, 3, 4, 23, 24, 29, 30, 34, 37, 39, 41, 44, 49

(Source: Analysis results from the authors' data, 2021)

The statistics in Table 3 show that among the five English 1 test papers that have been used from 2017 to 2021, three of them contained a large number of unsatisfactory questions such as: 18 items in the 2018 exam papers (accounting for 36%), 19 items in 2020 exam papers (accounting for 38%) and 14 items in 2021 exam papers (accounting for 28%). In addition, among the unsatisfactory items in the exam papers, some were unsatisfactorily chosen because of both difficulty and discrimination parameters. Specifically, item 43 in the 2017 paper; items 14, 17, 28, 35, 43, 48, 49 in the 2018 exam paper; items 4, 8, 32, 38, 40, 44 in the 2020 exam and the items 24, 30, 39, 41 in the 2021 exam paper. Among the unsatisfactory items on the exam papers, some had the abnormal values in terms of difficulty and discrimination. These items are presented in Table 4.

The statistics results in Table 4 show that the English 1 exam papers used in 2018, 2020 and 2021 included some items with very big or very small difficulty values. These items were obviously not meaningful in measuring the students' actual competence. In addition, some items in the above exam papers had negative discriminant values (a < 0.0). When they answered the items with negative discriminant value, high-performing students had a lower probability of giving the correct answers than low-performing ones. This is unreasonable for an objective multiple-choice item. Thus, the items in Table 4 seriously violated the requirements for multiple-choice items in the exam papers, so these items should be removed and should not be used in any exam papers.

Table 4. *Items of abnormal values in terms of difficulty and discriminant*

Exam paper	Item number	The value of parameters	
		Difficulty	***Discriminant***
2018	10	-1.95	-0.46
	35	-32.84	-0.03
	43	-51.50	-0.03
2020	4	-64.03	-0.02
	8	-16.46	-0.07
	32	251.46	0.00
	38	-4.31	-0.17
	41	-0.71	-0.59
2021	2	0.08	-0.10
	24	-75.34	-0.01
	30	-16.75	-0.05

(Source: Analysis results from the authors' data, 2021)

3.2. Evaluation of the English 1 exam papers by using the information curves

The information curve of the test will show the essential characteristics of the test as well as the level of students' competence that the test can measure accurately. Specifically, the maximum point of the curve with the horizontal axis is the level of competence that the test has the most accurate measurement meaning and the vertical axis is the level of information that the test provides. The results of the information curve display for English 1 exam papers over the school years are shown in Figure 1.

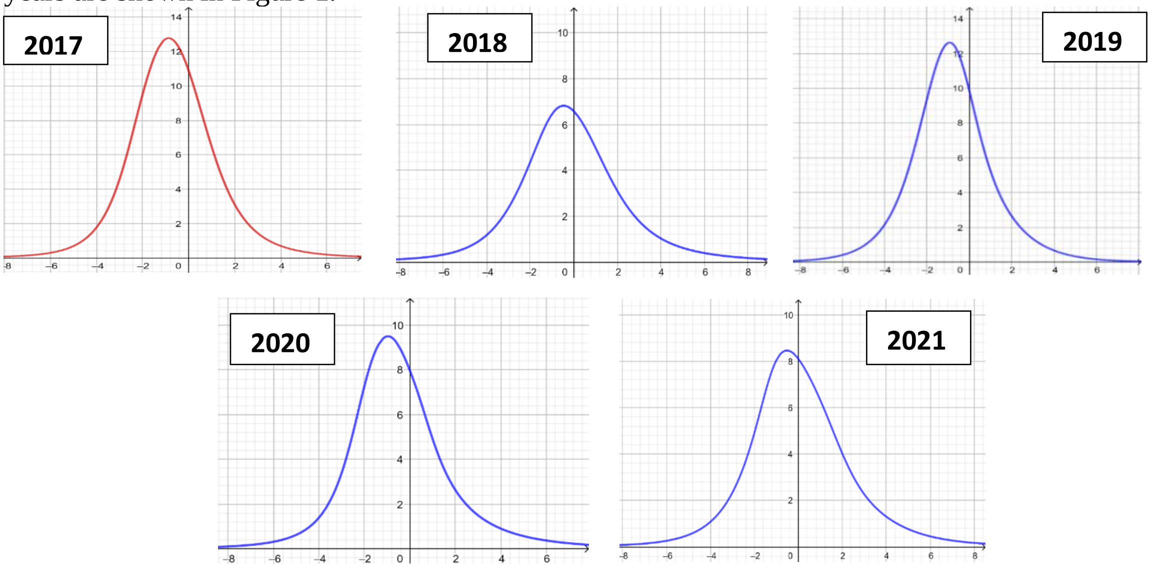


Figure 1. *The respective information curves for the English 1 papers from 2017 to 2021*

(Source: Analysis results from the authors' data, 2021)

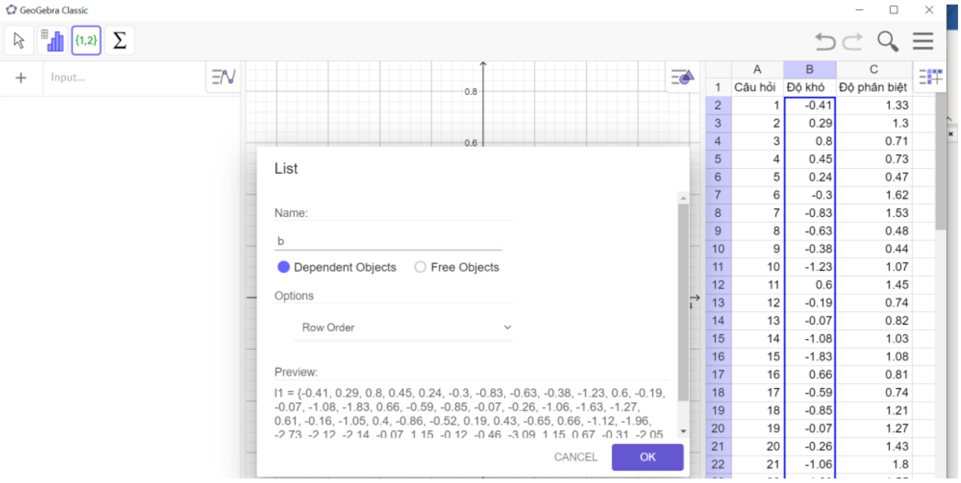
The displays of information curves of the English 1 test papers used from 2017 to 2021 in Figure 1 showed that the level of competence that the test questions were meaningfully measured is less than 0.0. Thus, it can be seen that the above test items accurately measured a relatively low level of competence. In addition, the information curves in the figures above show that the information level of the English 1 exam questions used over the years was not uniform. Specifically, the 2017 English 1 exam had the highest level of information at about 13.0, and the lowest was 2018 exam papers with information level below 7.0.

The above situation comes from the fact that the exam papers were written according to the lecturers' subjective experience while the items had not been analyzed and evaluated based on any scientific theories of education measurement namely IRT before being introduced into use, which has resulted in unsatisfactory items in terms of the difficulty and discrimination parameters. This situation will affect the work of testing and assessment for the wished aim of achieving the course's output standards. Of the same course, the level of information provided in the exam papers (tests) and the level of students' competence that the tests could measure were different.

To overcome the above situation, in addition to a thorough analysis, the tests should be evaluated by the information curve for the level of information they provide as well as the level of students' competence they will measure. To do this, it is proposed that GeoGebra should be used to determine the information curve for the multiple-choice items and test items based on the values of the parameters of difficulty, discrimination. In this study, the 2-parameter model was used to illustrate the information curve for the multiple-choice items. The expression of the information function of the multiple-choice test according to the above model was described in formula (6) (**2.3.1**). The process of drawing the information curve for the multiple-choice test by GeoGebra was done in by following the steps below:

Step 1. Input the characteristic parameters (*difficulty, discrimination*) of the item into GeoGebra. This can be done easily by some simple manipulations on the software interface as shown in Figure 2.

Step 2. Assign the discrimination parameter value of the item to the value *a*, the difficulty parameter of the item to the value *b*. These are the parameters used in the information curve expression of the multiple-choice test. To perform parameter value assignment of the items, software users must follow these steps: (1) Select all difficulty parameters of the items in the test (at the right part of Figure 2); (2) Select the icon {1,2} on the GeoGebra software interface; (3) Name the parameter as *b* in the Name box, then select OK as shown in Figure 2.



Name: b

● Dependent Objects ○ Free Objects

Options: Row Order

Preview:

l1 = {-0.41, 0.29, 0.8, 0.45, 0.24, -0.3, -0.83, -0.63, -0.38, -1.23, 0.6, -0.19, -0.07, -1.08, -1.83, 0.66, -0.59, -0.85, -0.07, -0.26, -1.06, -1.63, -1.27, 0.61, -0.16, -1.05, 0.4, -0.86, -0.52, 0.19, 0.43, -0.65, 0.66, -1.12, -1.96, -2.73, -2.12, -2.14, -0.07, 1.15, -0.12, -0.46, -3.09, 1.15, 0.67, -0.31, -2.05

	A	B	C
1	Câu hỏi	Độ khó	Độ phân biệt
2	1	-0.41	1.33
3	2	0.29	1.3
4	3	0.8	0.71
5	4	0.45	0.73
6	5	0.24	0.47
7	6	-0.3	1.62
8	7	-0.83	1.53
9	8	-0.63	0.48
10	9	-0.38	0.44
11	10	-1.23	1.07
12	11	0.6	1.45
13	12	-0.19	0.74
14	13	-0.07	0.82
15	14	-1.08	1.03
16	15	-1.83	1.08
17	16	0.66	0.81
18	17	-0.59	0.74
19	18	-0.85	1.21
20	19	-0.07	1.27
21	20	-0.26	1.43
22	21	-1.06	1.8

Figure 2. *Assigning the items' parameter values in to GeoGebra*

(Source: The authors' data, 2021)

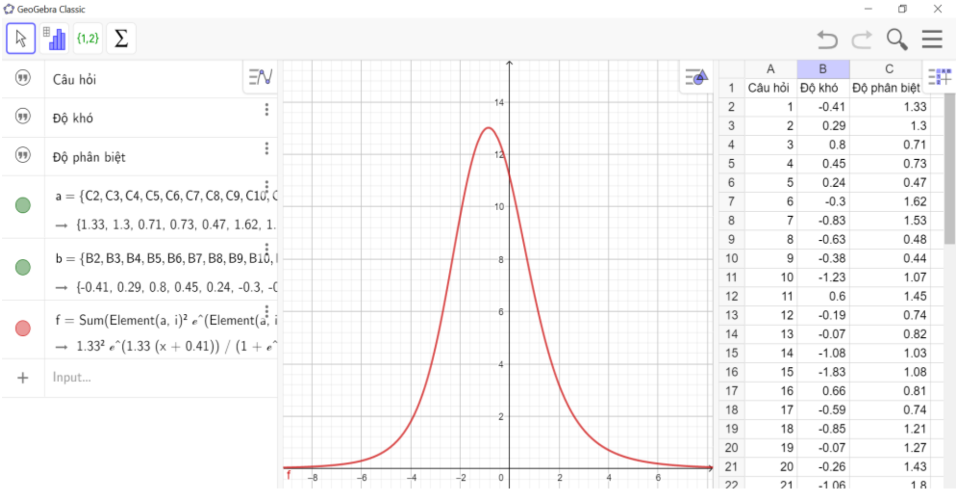
The result of parameter assignment is shown in Figure 3 at the left position.

Step 3. Input the command of the information curve for the multiple-choice test into the GeoGebra as the following code line:

```
Sum(Element(a, i)² e^(Element(a, i)(x-Element(b, i)))/(1+e^(Element(a, i)(x-Element(b, i))))², i, 1,
Length(b))
```

After entering the graphing command code line as above, the software will automatically draw and display the information curve for the multiple-choice test containing items with evaluated parameter properties.

The display of the information curve for the 2017 English 1 exam paper drawn by GeoGebra can be illustrated in Figure 3.



Câu hỏi

Độ khó

Độ phân biệt

a = {C2, C3, C4, C5, C6, C7, C8, C9, C10, C
→ {1.33, 1.3, 0.71, 0.73, 0.47, 1.62, 1.

b = {B2, B3, B4, B5, B6, B7, B8, B9, B10, I
→ {-0.41, 0.29, 0.8, 0.45, 0.24, -0.3, -C

f = Sum(Element(a, i)² e^(Element(a, i
→ 1.33² e^(1.33 (x + 0.41)) / (1 + e^

	A	B	C
1	Câu hỏi	Độ khó	Độ phân biệt
2	1	-0.41	1.33
3	2	0.29	1.3
4	3	0.8	0.71
5	4	0.45	0.73
6	5	0.24	0.47
7	6	-0.3	1.62
8	7	-0.83	1.53
9	8	-0.63	0.48
10	9	-0.38	0.44
11	10	-1.23	1.07
12	11	0.6	1.45
13	12	-0.19	0.74
14	13	-0.07	0.82
15	14	-1.08	1.03
16	15	-1.83	1.08
17	16	0.66	0.81
18	17	-0.59	0.74
19	18	-0.85	1.21
20	19	-0.07	1.27
21	20	-0.26	1.43
22	21	-1.06	1.8

Figure 3. *The display of information curve for the 2017 English 1 exam paper by GeoGebra*

(Source: Analysis results from the authors' data, 2021)

Based on how the information curve for the multiple-choice item displays, a test editor can evaluate the students' ability level that the item actually measures and the information that the item reveals before deciding to put it into official use. If an item is identified as inappropriate, the test editor can replace it by more suitable ones based on the content of knowledge and the values of the parameters of difficulty, their discriminant, and get it checked by GeoGebra (Step 1). At that time, the software will update the parameter values of the item and display the information curve. In this way, the test editor can write or select the appropriate test items, accurately assess the students' competence, and at the same time achieve the goal of the assessment.

4. Conclusion

This study has provided a method of analyzing and evaluating multiple-choice test items based on scientific theory of measurement in education, especially the analysis of test data to determine the characteristic parameters of each multiple-choice item, thereby identifying satisfactory items for selection or unsatisfactory for adjustment and improvement. This is especially useful for writing multiple-choice question banks for subjects in order to serve the work of assessment activities. In addition, the study also introduces the use of the information curve tool for a multiple-choice item and applies it to the design of multiple-choice items capable of accurately assessing students' competence. Research results have shown that the application of IRT in analyzing test items and writing multiple-choice tests are of very urgent and useful meanings.

REFERENCES

[1] T. Q. Lam, *Measurement in Education - Theory and Application*. Hanoi: Vietnam National University Press, Hanoi, 2011.

[2] G. Rasch, *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research, 1960.

[3] A. Birnbaum, "Some Latent Trait Models and Their Use in Inferring an Examinee's Ability," in *Statistical Theories of Mental Test Scores,* F. M. Lord and M. R. Novick, Eds., Addison-Wesley, 1968, pp. 397-479.

[4] F. Baker, *The basic of item response theory*. Washington: ERIC Clearinghouse on Assessment and Evaluation, 2001.

[5] S. L. V. Bortolotti, R. Tezza, D. F. de Andrade, A. C. Bornia, and A. F. de Sousa Júnior, "Relevance and advantages of using the item response theory," *Quality & Quantity,* vol. 47, no. 4, pp. 2341-2360, 2013.

[6] S. Rakkapao, S. Prasitpong, and K. Arayathanitkul, "Analysis test of understanding of vectors with the three-parameter logistic model of item response theory and item response curves technique," *Physical review physics education research,* vol. 12, no. 2, 2016, Art. no. 020135.

[7] C. B. Essen, I. E. Idaka, and M. A. Metibemu, "Item level diagnostics and model-data fit in item response theory (IRT) using BILOG-MG v3.0 and IRTPRO v3.0 programmes," *Global Journal of Educational Research,* vol. 16, no. 2, pp. 87-94, 2017.

[8] G. C. Foster, H. Min, and M. J. Zickar, "Review of item response theory practices in organizational research: Lessons learned and paths forward," *Organizational Research Methods,* vol. 20, no. 3, pp. 465-486, 2017.

[9] F. A. Setiawati, R. E. Izzaty, and V. Hidayat, "Items parameters of the space-relations subtest using item response theory," *Data in brief,* vol. 19, pp. 1785-1793, 2018.

[10] M. S. Mu'iz, I. Kaniawati, and T. R. Ramalis, "Analyzing instrument characteristics of critical thinking skills and mastery of concepts based on item response theory," *International Conference on Mathematics and Science Education of Universitas Pendidikan Indonesia,* vol. 3, pp. 162-167, 2018.

[11] T. Q. Lam, M. N. Lam, T. M. Le, and B. D. Vu, "VITESTA software and analysis of test data," (in Vietnamese), *Vietnam Journal of Education,* vol. 176, pp. 10-12, 2007.

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

[13] V. A. Le, U. H. Pham, C. H. Doan, and H. T. Le, "Using Gibbs Sampler to evaluate item difficulty in Rasch model," (in Vietnamese), *Ho Chi Minh City University of Education Journal of Science,* vol. 14, no. 4, pp. 119-130, 2017.

[14] K. A. Bui and P. N. Bui, "Using IATA to analyze, evaluate and improve the quality of the multiple-choice items in chapter power functions, exponential functions and logarithmic functions," (in Vietnamese), *Can Tho University Journal of Science,* vol. 54, no. 9C, pp. 81-93, 2018.

[15] C. V. Nguyen and H. P. Nguyen, "Analyzing and selecting multiple-choice test items based on Classical Test Theory and Item Response Theory," (in Vietnamese), *Ho Chi Minh city University of Education Journal of Science,* vol. 17, no. 10, pp. 1804-1818, 2020.

[16] C. H. Doan, V. A. Le, and U. H. Pham, "Applying three-parameter logistic model in validating the level of difficulty, discrimination and guessing of items in a multiple-choice test," (in Vietnamese), *Ho Chi Minh City University of Education Journal of Science,* vol. 7, no. 8, pp. 174-184, 2016.

[17] T. H. B. Nguyen, "Using Quest software to analyze objective test questions," (in Vietnamese), *Journal of Science and Technology - Da Nang University,* vol. 2, pp. 119-126, 2008.

[18] Q. N. Bui, "Evaluation of the quality of multiple-choice test bank for the module of Introduction to Anthropology by using the RASCH model and QUEST software," (in Vietnamese), *Science of Technology Development - Viet Nam National University Ho Chi Minh City,* vol. 20, no. X3, pp. 42-54, 2017.

[19] C. V. Nguyen and T. Q. Nguyen, "Applying ConQuest software with the two-parameter IRT model to evaluate the quality of multiple-choice test," (in Vietnamese), *HNUE Journal of Science,* vol. 65, no. 7, pp. 230-242, 2020.

[20] H. P. Nguyen and N. T. Du, "The analysis and selection of objective test items based on S-P chart, Grey Relational Analysis, and ROC curve," (in Vietnamese), *Ho Chi Minh City University of Education Journal of Science,* vol. 6, no. 72, pp. 163-173, 2015.

[21] H. P. Nguyen, "Using GSP chart and ROC method to analyze and select multiple-choice items," (in Vietnamese), *Dong Thap University Journal of Science,* vol. 24, no. 2, pp. 11-17, 2017.

[22] D. Rizopoulos, "An R package for latent variable modeling and item response theory analysis," *Journal of Statistical Software,* vol. 17, no. 5, pp. 1-25, 2006.

[23] M. Hohenwarter and J. Preiner, "Creating mathlets with open source tools," *The Journal of Online Mathematics and Its Applications,* vol. 7, pp. 1-29, 2007.