APPLICATION OF DIGITAL TRANSFORMATION IN EVALUATING MULTIPLE-CHOICE QUESTION BANKS

Nguyễn Thế Anh*, Hoàng Thị Nguyệt, Trần Anh Thảo

Học viện Hải quân

Thông tin chung:
Ngày nhận bài: 9/01/2025
Ngày phản biện: 12/01/2025
Ngày duyệt đăng: 04/3/2025
**Tác giả chính:*
nguyentheanh.uet@gmail.com
DOI:
https://doi.org/10.70879/Td3bU1A0y

Title:
Ứng dụng chuyển đổi số trong đánh giá các bộ ngân hàng câu hỏi trắc nghiệm

Từ khóa:
trắc nghiệm khách quan, chuyển đổi số, đánh giá chất lượng câu hỏi, lý thuyết trắc nghiệm cổ điển, lý thuyết ứng đáp câu hỏi.

Keywords:
objective multiple-choice, digital transformation, question quality assessment, Classical Test Theory, Item Response Theory.

TÓM TẮT*: Hiện nay, quá trình phân tích và đánh giá chất lượng câu hỏi trắc nghiệm khách quan yêu cầu giảng viên phải tự thống kê dữ liệu thi thủ công, trước khi sử dụng phần mềm để phân tích. Ngoài ra để biết cách sử dụng được các phần mềm này đòi hỏi người sử dụng phải được trang bị các kiến thức trong lĩnh vực đo lường và đánh giá. Điều này gây trở ngại không nhỏ đối với giảng viên. Bài báo này giới thiệu một giải pháp ứng dụng chuyển đổi số để thực hiện tự động hoàn toàn quy trình trên giúp giảng viên nhận được kết quả đánh giá nhanh chóng. Giải pháp mà nhóm tác giả đề xuất là xây dựng mô-đun bổ sung có chức năng thu thập dữ liệu thi trên hệ thống thi trắc nghiệm trực tuyến đang triển khai tại trường của chúng tôi. Sau đó xây dựng các thuật toán cần thiết để tính toán các tham số của câu hỏi từ dữ liệu đã thu thập. Cuối cùng đưa ra kết quả đánh giá dựa trên lý thuyết trắc nghiệm cổ điển và lý thuyết ứng đáp câu hỏi. Kết quả triển khai thực tế cho thấy hệ thống giảm đáng kể thời gian và công sức của giảng viên so với phương pháp thủ công, đồng thời nâng cao độ chính xác và tính khách quan trong quá trình đánh giá chất lượng câu hỏi.*

ABSTRACT*: At present, the process of analyzing and evaluating the quality of objective multiple-choice questions requires instructors to manually gather test data before being able to utilize software for analysis. Moreover, proficiency in the theory of measurement and evaluation is necessary to effectively use such software. This presents a significant challenge for instructors. This article introduces a digital tranformation solution aimed at completely automating this process, providing instructors with prompt evaluation results. The proposed solution involves the development of an additional module that collects test data from an online multiple-choice examination system currently being implemented at our institution. Subsequently, the necessary algorithms are implemented to calculate question parameters from the collected data. Finally, evaluation results are generated based on classical test theory and item response theory. Real-world implementation demonstrates that the system significantly reduces instructors' time and effort compared to manual methods, while enhancing the accuracy and objectivity of question quality assessment.*

1. Introduction

In recent years, objective multiple-choice questions have become a widely adopted tool in higher education institutions for assessing learners' competencies, owing to their objectivity and efficiency in evaluating broad knowledge domains [7]. However, ensuring the effectiveness of this method requires rigorous analysis and evaluation of question quality prior to implementation. Such evaluations typically draw on Classical Test Theory (CTT) and Item Response Theory (IRT) [1, p.19]. Previous research has employed various tools to assess multiple-choice question quality. For example, Pham and Nguyen (2021) and Rizopoulos (2017) utilized the "ltm" package in R for IRT-based data analysis [2], [3], while Lam et al. (2007) developed the VITESTA software for question evaluation [4]. Similarly, Bui et al. (2018) applied IATA to enhance question banks [5], and Nguyen (2008) used QUEST to optimize question quality [6]. Despite these advancements, these approaches often rely on manual data compilation and demand specialized knowledge of measurement theories, posing significant challenges for instructors, including time-intensive processes and the potential for errors in data handling.

To address these challenges, automating the analysis of test data has become a pressing need. The traditional process involves collecting statistics on question counts, test-taker numbers, and response patterns [1], followed by software-based computation of parameters such as difficulty and discrimination indices [8]. Yet, the large data volumes involved make this approach cumbersome and error-prone. Building on this context, the Naval Academy has embraced digital transformation by implementing an online multiple-choice testing system, which facilitates automated data collection. To overcome the limitations of manual methods, we have enhanced this system by integrating CTT and IRT algorithms, enabling full automation of multiple-choice question data analysis. This solution significantly reduces instructors' time and effort, ensures accuracy and objectivity, and strengthens question bank quality, ultimately enhancing learner competency assessment [9]. Such digital transformation is pivotal in modernizing education and meeting the demands of innovative training institutions.

1. Research methodology

2.1. Overview of CTT and IRT

Classical Test Theory (CTT) and Item Response Theory (IRT) are two important theories in the field of measurement and assessment of tests [1] (p.19). Both theories are built upon probability and statistical principles.

Classical Test Theory (CTT) treats a test as a set of independent questions, focusing on two key parameters: difficulty (the proportion of test-takers who answer a question correctly) and discrimination (the ability of a question to differentiate between high- and low-ability test-takers). These parameters evaluate question and test quality but do not account for individual test-taker characteristics, limiting their ability to model nuanced response patterns.

Item Response Theory (IRT), in contrast, provides a more sophisticated approach by modeling the relationship between a test-taker's ability and their probability of answering a question correctly. IRT estimates both question parameters and test-taker abilities based on response patterns, offering detailed insights into question quality and individual performance. IRT employs several models, with the one-parameter logistic (1PL), two-parameter logistic (2PL), and three-parameter logistic (3PL) models being the most common, each defined by specific parameters:

2.1.1.One-Parameter Logistic Model (1PL, Rasch Model): This model uses a single parameter, *difficulty* (denoted as *b*). The 1PL model assumes all questions have the same discrimination ability, simplifying analysis but limiting its flexibility. The probability of a correct response is given by: $P(\theta) = \frac{1}{1+e^{-(\theta-b)}}$ where (θ) is the test-taker's ability, and b is the question's difficulty.

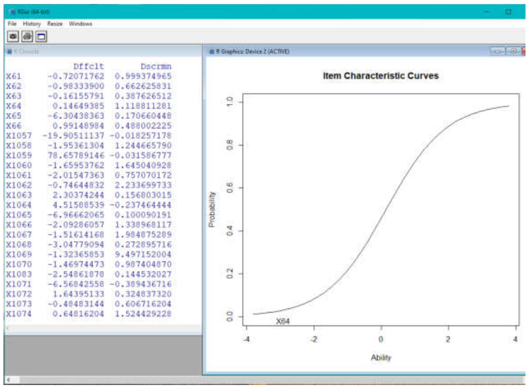
2.1.2.Two-Parameter Logistic Model

(2PL): The 2PL model extends the 1PL by adding a *discrimination* parameter (denoted as *a*). A higher *a* value indicates a steeper item characteristic curve, meaning the question is more effective at distinguishing between test-takers of similar ability levels. The probability of a correct response is: $P(\theta) = \frac{1}{1+e^{-a(\theta-b)}}$. In this model, *a* and *b* are estimated for each question, providing a more nuanced evaluation of question quality. For example, a question with high discrimination *a* and moderate difficulty *b* is ideal for distinguishing between test-takers of varying abilities.

2.1.3.Three-Parameter Logistic Model (3PL): The 3PL model includes an additional *guessing* parameter (denoted as *c*), which accounts for the probability that a low-ability test-taker answers correctly by guessing. This is particularly relevant for multiple-choice questions, where random guessing can inflate correct response rates. The probability of a correct response is: $P(\theta) = c + (1-c)\frac{1}{1+e^{-a(\theta-b)}}$ The guessing parameter *c* typically ranges from 0 to 0.25 for four-option multiple-choice questions, reflecting the chance of guessing correctly. The 3PL model is more complex but provides a realistic representation of test-taking behavior, especially in high-stakes exams.

These IRT parameters are estimated using statistical methods, such as Maximum Likelihood Estimation or Bayesian approaches, implemented in software like R's "ltm" package [3]. The estimated parameters are used to generate *item characteristic curves* (ICCs), which plot the probability of a correct response against test-taker ability. For instance, a question with low *b*, high *a*, and low *c* produces a steep ICC centered at a lower ability level, indicating an easy but highly discriminative question. Conversely, a question with high *b*, low *a*, or high *c* may be problematic, as it fails to accurately assess ability or is susceptible to guessing.

IRT's ability to model individual test-taker abilities and question characteristics makes it a modern, advanced method for evaluating multiple-choice questions. By analyzing parameters like *a*, *b*, and *c*, instructors can identify questions that are too easy, too difficult, poorly discriminative, or prone to guessing, enabling targeted improvements to test design. Manual calculation of CTT and IRT parameters is labor-intensive and error-prone. Modern software, such as R, VITESTA, and IATA [2], automates these calculations, enhancing accuracy and efficiency. For example, the "ltm" package in R estimates IRT parameters and generates ICCs, allowing instructors to visually assess question quality without extensive statistical expertise.



```
            Dffclt        Dscrmn
X61     -0.72071762  0.999374965
X62     -0.98333900  0.662625831
X63     -0.16155791  0.387626512
X64      0.14649385  1.118811281
X65     -6.30438363  0.170660448
X66      0.99148984  0.488002225
X1057  -19.90511137 -0.018257178
X1058   -1.95361304  1.244665790
X1059   78.65789146 -0.031586777
X1060   -1.65953762  1.645040928
X1061   -2.01547363  0.757070172
X1062   -0.74644832  2.233699733
X1063    2.30374244  0.156803015
X1064    4.51588539 -0.237464444
X1065   -6.96662065  0.100090191
X1066   -2.09286057  1.338968117
X1067   -1.51614168  1.984875289
X1068   -3.04779094  0.272895716
X1069   -1.32365853  9.497152004
X1070   -1.46974473  0.987404870
X1083   -2.54861878  0.144532027
X1071   -6.56842558 -0.389436716
X1072    1.64395133  0.324837320
X1073   -0.48483144  0.606716204
X1074    0.64816204  1.524429228
```

Figure 1. *Computed item parameters for a multiple-choice test using IRT (2PL model), obtained using the R software.*

In Figure 1, the item parameters, including difficulty (abbreviated as "Dffclt") and discrimination (abbreviated as "Dscrmn"), are represented in two columns. The rows correspond to the coded questions, such as X61, X62, etc. Using these two parameters, the software can plot the item characteristic curve (e.g., the graph for question X64 as shown in the figure). This visual representation allows instructors to assess the questions intuitively. By examining the ordinate values of the graph at the abscissa of

0, one can determine whether the question is easy or difficult. The closer the value is to 1, the easier the question. Furthermore, a higher gradient of the graph indicates better discriminative of the question in assessing the abilities of the test takers [1]. On the other hand, problematic questions often have ordinate values close to 0 or 1 (too difficult or too easy), a low gradient, or even a negative gradient (downward graph).

2.2. Process of Analyzing Objective Multiple-Choice Test Data

The process of analyzing data from objective multiple-choice tests typically involves the following steps:

Step 1: Data Preparation

The first step in the process of analyzing data from objective multiple-choice tests is data preparation. This step includes collecting data from the objective multiple-choice test and ensuring that the data is accurately and completely entered. It is important to ensure that the data collection process is carried out accurately and supervised.

Step 2: Data Preprocessing

Data preprocessing is the process of preparing and cleaning the data before performing analyses and modeling. The goal of this step is to make the data accurate, consistent, and easy to handle such as checking and removing invalid responses, handling missing or duplicate responses by replacing or eliminating them.

Data preprocessing is an important part of the data analysis process, enhancing the accuracy and reliability of the analysis results.

Step 3: Data Analysis

Data analysis is the process of applying methods and techniques from CTT and IRT to calculate the characteristic parameters of the questions, such as difficulty and discrimination (according to CTT), or estimating the parameters (according to IRT), and assessing the appropriateness of the IRT model with the test data.

Currently, data analysis is facilitated by software tools. For example, the free software R, with its "ltm" package [4], provides tools for analyzing data from tests and evaluating them using various IRT models.

Step 4: Reporting the Results

Presenting and conveying the results of data analysis from software in a clear and understandable manner. For example, by grouping questions: a group of questions with good parameters (within an acceptable range) and a group of questions with bad parameters (parameters outside the acceptable range). Utilize the results, specifically the calculated or estimated parameters, to draw meaningful conclusions: identifying questions with good quality, highlighting questions that need further consideration or adjustment, and even considering their removal if necessary.

2.3. The Necessity of Digital Transformation in Objective Multiple-Choice Test Data Analysis

In the process of analyzing data from objective multiple-choice tests, the data preparation and preprocessing steps are often the most time-consuming and labor-intensive stages. Specifically, objective multiple-choice test data tends to have a large scale with a significant number of questions. Moreover, to ensure objectivity in the test administration process, the answer choices are often randomized. As a result, the data collection process requires careful attention to detail and precision.

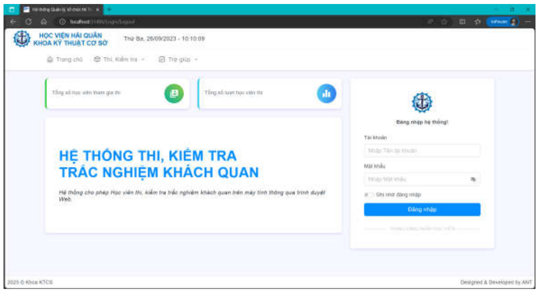
In the case of paper-based objective multiple-choice tests, data collection is even more complex and time-consuming. To address this issue, an important requirement is to transform the test format from paper-based to online objective multiple-choice testing. This helps save time and effort in the data collection process. When using online objective multiple-choice testing, the test results are automatically collected, and the data is automatically processed. Test-takers complete the test on an online interface, and the system automatically records their answers. This process not only reduces effort and time but also ensures the accuracy and consistency of the data.

However, if the transition is simply from paper-based to online testing, there is still a barrier that users need to overcome, which is the requirement to know how to use the test data analysis software. Additionally, a clear understanding of CTT and IRT is necessary in order to draw conclusions about the parameters that the software has calculated or estimated to evaluate the quality of the objective multiple-choice test items.

Therefore, digital transformation of the entire process of analyzing data from objective multiple-choice tests becomes necessary. This not only helps save time and effort for instructors in evaluating the quality of the questions but also ensures accuracy in assessing learners' abilities, thereby enabling the implementation of appropriate teaching strategies.

2.4. Digital Transformation Solution for Analyzing Objective Multiple-Choice Test Data

Currently, the Naval Academy has implemented a digital transformation solution for the testing and assessment process by developing an online objective multiple-choice test system (Figure 2). This system is used to assess learners' abilities through objective multiple-choice test evaluations in various subjects. However, we have recognized the potential to leverage the statistical capabilities of this system to enhance the quality of the evaluation process.



HỌC VIỆN HẢI QUÂN
KHOA KỸ THUẬT CƠ SỞ

Thứ Ba, 26/09/2023 - 10:10:09

HỆ THỐNG THI, KIỂM TRA
TRẮC NGHIỆM KHÁCH QUAN

Hệ thống cho phép Học viên thi, kiểm tra trắc nghiệm khách quan trên máy tính thông qua trình duyệt Web.

Đăng nhập hệ thống!

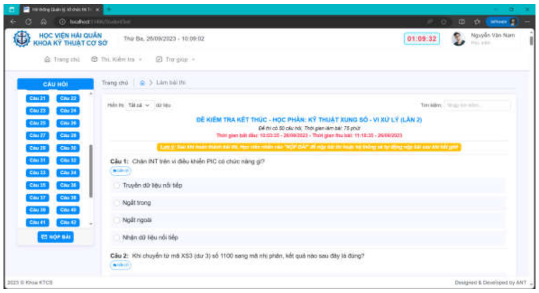
Tài khoản

Mật khẩu

2023 © Khoa KTCS

Designed & Developed by ANT

(a)



HỌC VIỆN HẢI QUÂN
KHOA KỸ THUẬT CƠ SỞ

Thứ Ba, 26/09/2023 - 10:09:02

01:09:32

Nguyễn Văn Nam

ĐỀ KIỂM TRA KẾT THÚC - HỌC PHẦN: KỸ THUẬT XUNG SỐ - VI XỬ LÝ (LẦN 2)

Đề thi có 50 câu hỏi; Thời gian làm bài: 75 phút

Câu 1: Chân INT trên vi điều khiển PIC có chức năng gì?

○ Truyền dữ liệu nối tiếp

○ Ngắt trong

○ Ngắt ngoài

○ Nhận dữ liệu nối tiếp

Câu 2: Khi chuyển từ mã XS3 (dư 3) số 1100 sang mã nhị phân, kết quả nào sau đây là đúng?

2023 © Khoa KTCS

Designed & Developed by ANT

(b)

Figure 2. *Online Objective Multiple-Choice Testing System being implemented at the Naval Academy: (a) login page and (b) exam page*

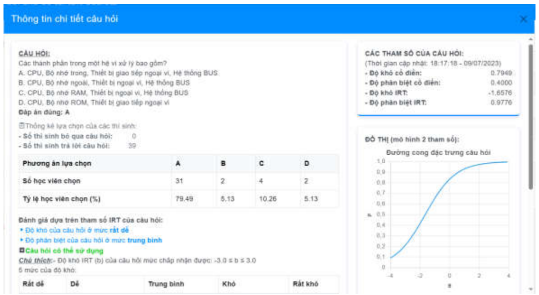
For the system we are using, the test data is collected and stored in a server database. However, this data is only utilized for scoring the exams and displaying the individual test results for each student. Therefore, the research question arises as to how we can use this data for the purpose of analyzing and evaluating the quality of the questions. As a result, we have continued to upgrade the system to fully automate the process of analyzing the objective multiple-choice test data. This solution incorporates two important theories in the field of educational measurement, namely Classical Test Theory (CTT) and Item Response Theory (IRT), to achieve the goal of optimizing the analysis and evaluation of question quality. To address this issue, we collect additional data for each question in the exam, including the students' responses to the questions (both correct and incorrect), data on students skipping questions, and data on the number of students selecting each answer option. These are the necessary input data for calculating the question parameters based on CTT and IRT.

For instance, according to CTT, the difficulty index of a multiple-choice question is defined as the percentage of students who answered it correctly out of the total number of participating students [1, p.59]. Therefore, by fully collecting and analyzing these two statistics, we can determine the difficulty index of the question. Similarly, the discrimination index of a question is also calculated based on the data of students who answered it correctly within the high-performing and low-performing groups, using the formula provided by CTT [1, p.61]. In the case of IRT, difficulty and discrimination are estimated by employing parameter estimation methods for multiple-choice questions [1, p.102]. Developing algorithms based on IRT can be more complex compared to CTT. Hence, we have chosen to apply the algorithm from the "ltm" package [4] in the R software. This package is an open-source extension, freely available, highly regarded by experts in statistical analysis and widely used in the field of data assessment.

Once the parameters of the questions have been determined, we proceed to integrate the algorithms into the system based on the test theories in order to assess the difficulty level of the questions on a 5-point scale (very easy, easy, moderate, difficult, very difficult) and the discrimination of the questions on a 5-point scale (very poor, poor, moderate, good, very good) [5]. Consequently, the system automatically generates conclusions about which questions can be used and which questions need review, allowing the test creator to have an objective assessment of the questions without requiring expertise in test theories. Additionally, the authors have constructed item characteristic curve graphs and item information curve graphs to quickly identify problematic questions for users with expertise in the field. Thus, the entire process of analyzing and evaluating question quality is fully automated by the system, from organizing and statistically analyzing the test data to performing calculations and generating conclusions.

3. Results and discussion

By applying Classical Test Theory (CTT) and Item Response Theory (IRT), the digital transformation solution for analyzing objective multiple-choice test data has established an effective and reliable approach to assessing question quality. The process involves collecting test data and preprocessing it using tailored algorithms for input into the system. Subsequently, data analysis algorithms are utilized to calculate and estimate parameters of question difficulty and discrimination. These results enable the system to provide a preliminary assessment of the questions, assisting test-makers in gaining a clear understanding of question quality for selection, review, or adjustment before implementation.



Thông tin chi tiết câu hỏi

CÂU HỎI:
Các thành phần trong một hệ vi xử lý bao gồm?
A. CPU, Bộ nhớ trong, Thiết bị giao tiếp ngoại vi, Hệ thống BUS
B. CPU, Bộ nhớ ngoài, Thiết bị ngoại vi, Hệ thống BUS
C. CPU, Bộ nhớ RAM, Thiết bị ngoại vi, Hệ thống BUS
D. CPU, Bộ nhớ ROM, Thiết bị giao tiếp ngoại vi
Đáp án đúng: A

Thống kê lựa chọn của các thí sinh:
- Số thí sinh bỏ qua câu hỏi: 0
- Số thí sinh trả lời câu hỏi: 39

Phương án lựa chọn	A	B	C	D
Số học viên chọn	31	2	4	2
Tỷ lệ học viên chọn (%)	79.49	5.13	10.26	5.13

Đánh giá dựa trên tham số IRT của câu hỏi:
- Độ khó của câu hỏi ở mức rất dễ
- Độ phân biệt của câu hỏi ở mức trung bình

Câu hỏi có thể sử dụng
Chú thích:- Độ khó IRT (b) của câu hỏi mức chấp nhận được: -3.0 ≤ b ≤ 3.0
5 mức của độ khó:

Rất dễ	Dễ	Trung bình	Khó	Rất khó

CÁC THAM SỐ CỦA CÂU HỎI:
(Thời gian cập nhật: 18:17:18 - 09/07/2023)

- Độ khó cổ điển:	0.7949
- Độ phân biệt cổ điển:	0.4000
- Độ khó IRT:	-1.6576
- Độ phân biệt IRT:	0.9776

ĐỒ THỊ (mô hình 2 tham số):
Đường cong đặc trưng câu hỏi

Figure 3. *Analysis results of a objective multiple-choice test by the system.*

The analysis results presented in Figure 3 offer a detailed evaluation of a specific question. According to CTT, the question has a difficulty index (P) of 0.7949, indicating it is easy, and a discrimination index (D) of 0.4000, suggesting moderate discrimination ability. Using IRT (2PL model), the difficulty parameter (b) is -1.6576, confirming the question as very easy, and the discrimination parameter (a) is 0.9776, indicating acceptable but not optimal discrimination [8]. The Item Characteristic Curve (ICC) in Figure 3 shows that test-takers with average ability (0) have an approximately 80% probability of answering correctly, reinforcing the question's ease. However, the moderate discrimination suggests that while the question effectively identifies high-ability test-takers, it may need adjustment to better distinguish between those with closer ability levels.

This solution brings numerous benefits. By providing precise parameters and visual tools like the ICC, the system enables instructors to accurately assess learner competencies and refine question banks effectively. It simplifies the identification of high-quality questions and those requiring revision, allowing for the design of more reliable examinations. Consequently, this enhances the accuracy of assessments and contributes to improving teaching quality in higher education institutions.

4. Conclusion

In the effort to enhance the accuracy of assessing learners' abilities, the application of digital transformation solutions in the analysis of data from multiple-choice tests to evaluate the quality of questions based on CTT and IRT has yielded noteworthy results. This solution enables a fully automated and efficient process of analyzing test data that is reliable. The utilization of technology and data analysis software optimizes the workflow, saves time, and ensures accuracy during the analysis process. This enables test creators to have a more objective view of the questions, allowing them to actively select questions with appropriate difficulty and discrimination parameters for inclusion in exams. This, in turn, contributes to accurately assessing learners' abilities and improving the quality of education and training in higher education institutions.

References

1. Q. T. Lam. (2010). *Measurement in Education: Theory and Application*. Hanoi, Vietnam: Vietnam National University, Hanoi.
2. V. T. Pham & V. C. Nguyen (2021). *Analysis and evaluation of multiple-choice test items and test design: A study on application of Item Response Theory*. TNU Journal of Science and Technology, 226(13), 72-81.
3. Rizopoulos, D. (2017). *ltm: An R package for latent variable modeling and item response theory analyses (Version 1.1-0).* Retrieved from https://cran.r-project.org/web/packages/ltm/index.html

4.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

5. K. A. Bui and P. N. Bui, *"Using IATA to analyze, evaluate and improve the quality of the multiplechoice 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.

6.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.

7.T.T. Duong (2005). *Objective Testing and Educational Achievement Measurement*. Hanoi, Vietnam: Social Science Publishing House, Hanoi.

8.Baker, F. B., (2001), *The basic of item response theory*, College Park, MD: University of Maryland, ERIC Clearinghouse on Assessment and Evaluation.

9. Long, V. D., Dung, N. V., Thao, V. T., & Linh, N. T. M., *“Calculation and comparison of the item difficulty based on classical test theory and item response theory by CETA/R programs”* (in Vietnamese), Vietnam Journal of Educational Sciences, vol. 36, pp. 13-18, 2020.