# ANALYSIS AND EVALUATION OF QUESTION ITEMS: A SOLUTION TO ENHANCE THE QUALITY OF MULTIPLE-CHOICE TEST

**Nguyen Van Canh[1*], Pham Van Tac[2], and Nguyen Quoc Tuan[1]**

*[1]Quality Assurance Office, Dong Thap University*

*[2]Foreign Language Faculty, Dong Thap University*

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

**Article history**

*Received: 26/9/2022; Received in revised form: 07/02/2023; Accepted: 17/02/2023*

**Abstract**

*Objective tests are among the most used assessment forms in educational institutions. However, designing multiple-choice tests of good quality is usually very difficult as it requires test designers to implement the testing, analysis and evaluation of question items for adjustment and improvement prior to use. This study presents how to analyze and evaluate multiple-choice questions based on Classical Test Theory. The data used in this study are the exam results performed by regular students majoring in Informatics Teacher education and Computer Science in their four basic Informatics exam papers in Dong Thap University, from the academic year of 2017-2018 to that of 2020-2021. Based on the parameters of the questions entirely calculated by Microsoft Excel software, the authors show how to classify good question items in the exam papers that can be included in the question banks for future use of testing and assessment activities, and at the same time how to identify unsatisfactory questions that should be revised for adjustment, improvement, or elimination.*

**Keywords:** *Difficulty index, discrimination index, Classical Test Theory, multiple - choice question, test analysis.*

DOI: https://doi.org/10.52714/dthu.12.3.2023.1042

Cite: Nguyen Van Canh, Pham Van Tac, and Nguyen Quoc Tuan. (2023). Analysis and evaluation of question items: A solution to enhance the quality of multiple-choice test. *Dong Thap University Journal of Science, 12*(3), 18-28.

# PHÂN TÍCH, ĐÁNH GIÁ CÂU HỎI: MỘT GIẢI PHÁP NÂNG CAO CHẤT LƯỢNG ĐỀ THI TRẮC NGHIỆM KHÁCH QUAN

**Nguyễn Văn Cảnh[1*], Phạm Văn Tặc[2] và Nguyễn Quốc Tuấn[1]**

*[1]Phòng Bảo đảm chất lượng, Trường Đại học Đồng Tháp*

*[2]Khoa Ngoại ngữ, Trường Đại học Đồng Tháp*

*[*]Tác giả liên hệ: nvcanh@dthu.edu.vn*

**Lịch sử bài báo**

*Ngày nhận: 26/9/2022; Ngày nhận chỉnh sửa: 07/02/2023; Ngày duyệt đăng: 17/02/2023*

**Tóm tắt**

*Trắc nghiệm khách quan là một trong những hình thức đánh giá đang được sử dụng khá phổ biến hiện nay trong các cơ sở giáo dục. Tuy nhiên, việc thiết kế được các đề thi trắc nghiệm khách quan có chất lượng tốt thường rất khó khăn, đòi hỏi người ra đề cần phải thực hiện việc thử nghiệm, phân tích và đánh giá các câu hỏi trước để điều chỉnh, cải tiến trước khi đưa vào sử dụng chính thức. Nghiên cứu này trình bày cách phân tích, đánh giá các đề thi trắc nghiệm khách quan dựa trên lý thuyết trắc nghiệm cổ điển. Dữ liệu được sử dụng trong nghiên cứu này là kết quả thi của sinh viên hệ chính quy chuyên ngành Sư phạm Tin học và Khoa học máy tính đối với 04 đề thi Tin học căn bản được sử dụng tại Trường Đại học Đồng Tháp, từ năm học 2017-2018 đến năm học 2020-2021. Dựa trên các tham số của các câu hỏi được tính toán hoàn toàn bằng phần mềm Microsoft Excel, các tác giả đã chỉ ra cách phân loại những câu hỏi tốt trong các đề thi có thể đưa vào ngân hàng câu hỏi để sử dụng cho việc kiểm tra đánh giá, đồng thời chỉ ra cách xác định những câu hỏi chưa đạt yêu cầu cần phải được xem xét lại để điều chỉnh, cải tiến hoặc loại bỏ.*

**Từ khóa:** *Câu hỏi trắc nghiệm, độ khó, độ phân biệt, lý thuyết trắc nghiệm cổ điển, phân tích đề thi.*

## 1. Introduction

Assessment is an integral phase in the teaching process, so assessment tools must be objective, reliable and accurately reflect the different levels of learners' achievement (Kheyami *et al.*, 2018). Among the assessment tools, a test is considered a measuring device intended to numerically describe a learner's level or load of learning. Therefore, it is important to evaluate the quality of the test items via appropriate measurement methods in order to identify their reliability. More specifically, the evaluation of question items in the tests is one way of evaluating their constituent elements, from which the validity of the question items can be revealed (Haladyna, 2004). Currently, multiple choice question items are being widely used in higher education as a means of complementation or even replacement of other assessment methods. The development of this assessment method has been driven by common changes in higher education environments such as an increase in student size, the need to reduce resources, changes in testing and assessment models and increasing availability of computer networks (Nicol, 2007). For certain limitations in this assessment method, many researchers have discouraged the use of objective tests because they promote memorization but discourage (or test) high cognitive processes (Airasian, 1994; Scouller, 1998). However, some other education researchers argue that this depends on how the tests are designed and that they can be used to assess learning at higher cognitive levels (Cox, 1976; Johnstone and Ambusaidi, 2000). In fact, writing a high-quality objective test is difficult, time-consuming, but the tests have their advantages in terms of their high objectivity while they remove test designers' partiality because the learners' responses can be easily and reliably graded, especially when assessment is done on a large number of test takers (Cronbach and Shavelson, 2004). In addition, objective tests help assess learners' large amount of knowledge objectively in a short time (Patil *et al.*, 2016). Furthermore, if designed correctly and scientifically, test question items help assess the leaners' level of understanding and application of knowledge and problem-solving skills (Al-Wardy, 2010).

One of the major challenges in using objective tests in the assessment process is how to successfully design high-quality question items. In particular, the question items must be tested, analyzed and evaluated before being introduced into official use (Odukoya *et al.*, 2018). This includes the process of collecting, synthesizing and using information from learners' responses to evaluate the quality of question items (Ary *et al.*, 2002; Carroll, 1993; Fowell *et al.*, 1999). Besides, the evaluation and analysis work reveal information about whether a question item is reliable and valid. On that basis, good question items will be identified and then be introduced into question banks, and unsatisfactory question items will need improving or eliminating (Considine *et al.*, 2005). One of the most widely approach to evaluate the quality of a test item is Classical Test Theory (Davies, 1990; Zubairi and Kassim, 2006), in which indicators to be concerned and put under consideration are the difficulty level and discrimination level of the question items (Zubairi and Kassim, 2006). This shows questions on a test paper are effective so as to differentiate high-performing examinees from low-performing candidates. In particular, question items with poor discrimination level should be revised to detect possible limitations (Bachman, 1990). Meanwhile, the higher the question discrimination level becomes, the better it is for assessment value. In addition, another scientific theory that is being used commonly in the analysis of objective tests is the Item Response Theory. For the Classical Test Theory, the unit of analysis is the tests, while for the Item Response Theory the unit of analysis is the questions (Baker, 2001; Hambleton *et al.*, 1991). In fact, among the defenders of each theory, there has been a debate about which theoretical aspect is better (Haladyna, 2004). Those who are in favor of Item Response Theory argue that the main limitation of Classical Test Theory is the impossibility to separate examiners' characteristics from those of the test. That is, it is impossible to compare the different examinees in terms of their performance when they answer different tests (Hambleton *et al.*, 1991). In addition, the question parameters that are calculated based on the used sample can be seen as a limitation. For example, the difficulty level for the same question may be higher or lower when the individuals in the sample have a higher or lower ability (Haladyna,

2004). However, the analysis of questions using Classical Test Theory is more intuitive and easier to perform, especially for those who have not been trained in measurement and assessment knowledge in education. Meanwhile, limitations of Item Response Theory have been raised regarding the size and heterogeneity of the used samples. In particular, if the samples are small and heterogeneous, then the parameter values of the calculated questions cannot be considered good estimators (Haladyna, 2004). Therefore, it is obvious that analysis and evaluation of question items play a very essential and indispensable role in designing question items. However, this job has not been very popular in educational institutions, so the quality of objective tests is of poor quality and fail to assess learners' ability correctly.

In this study, the authors analyze and evaluate objective tests by using classical multiple-choice theory in order to sort out good question items that can be then introduced to question banks, especially point out unsatisfactory question items that need revising for accuracy and improvement. Classical multiple-choice theory was used for the analysis and evaluation of the question items because the estimation of the parameters is very intuitive and easy to implement. The operations to calculate the parameters of the questions can be done entirely using basic functions in Microsoft Excel. This will facilitate lecturers to get to know the test design and apply them in analyzing questions for quality question items thereby accurately assessing students' abilities, contributing to improving the quality of students, and teaching and learning activities in universities.

## 2. Theoretical background and research methodology

### 2.1. Classical Test Theory

Classical Test Theory was found around the end of the nineteenth century, and it was finally completed in the 1960s. This theory is built based on statistical science and is mainly applied in the analysis and evaluation of objective test items. The evaluation of objective test items in accordance with the theory is mainly based on the parameters of difficulty level, discrimination level and correlation coefficient of the question items in comparison with the test after the test takers' feedback on the test question items is attained (Lam Quang Thiep, 2011).

***Difficulty level of question items***

The difficulty level (P) of a question item is the proportion calculated by the total number of examinees giving the correct answers over the total number of candidates. Thus, the smaller the P value gets, the higher the difficulty of the question becomes and vice versa. The difficulty level of an objective question item is acceptable when the P value is between 0.25 and 0.75, corresponding to the number of examinees who answer correctly from 25% to 75%. The question item is considered too easy when the P-value is > 0.75 (over 75% of the candidates answer correctly), while the question is considered too difficult when the P-value is < 0.25. In addition, for an objective test item with n options, the difficulty level of the question is average when $P = (1+1/n)/2$. Specifically, the questions with 4 options have an average difficulty level of P = 0.65 (corresponding to 65% of examinees answering correctly), questions with 5 options have an average difficulty level of P = 0.6 (corresponding to 60% of candidates with correct answers). According to Lam Quang Thiep (2011), a good multiple-choice test usually contains many question items of average difficulty level. Meanwhile, question items that are too easy with the difficulty level of P > 0.75 (corresponding to more than 75% of examinees with correct answers) or those too difficult with a difficulty level less than 0.25 (corresponding to less than 25% of the examinees with correct answers) would be considered for adjustment, improvement, or removal from the test.

***Discrimination level of question items***

Discrimination level of an objective test item is the question item's likeliness to make a distinction between a group of high-performing candidates and a group of low-performing examinees who will answer the question item itself. A question item with good discrimination is the one to which the group of high-qualified candidates must have a higher rate of receiving correct answer than that of the group of low-qualified examinees. In which, the group of candidates with high ability is 27% of the total number of examinees with high scores from top to bottom; group of candidates with low ability is 27% of the total number of examinees with low scores from bottom to top (Lam Quang Thiep, 2011). The discrimination level of the question is determined by the following formula:

$$D = \frac{N_c - N_t}{N} \qquad (1)$$

in which, D is the discrimination level of the question item, $N_c$ is the number of examinees in the high ability group who correctly completed the question item, $N_t$ is the number of candidates in the low ability group who correctly answered the question, N is 27% of the total number of examinees.

The discrimination level of the question item is divided into the following levels: very good when D≥0.4, quite good when 0.30≤D≤0.39, average when 0.20≤D≤0.29, and poor when D≤0.19 (Duong Thieu Tong, 2005; Ebel, 1972). Therefore, the question items used in the test should have a discrimination value of 0.2 or higher (Lam Quang Thiep, 2011).

***Correlation coefficient between the question items and the test***

The scores of the question items on the test should be correlated with the scores of the whole test. This value of correlation coefficient is determined according to the following formula:

$$r = \frac{(\bar{x}_i - \bar{x}_c)}{\sigma}\sqrt{\frac{p_i}{1-p_i}} \qquad (2)$$

With $\bar{x}_i$ being the average score of those who correctly answered the i-th question item considering the correlation with the multiple-choice test; $\bar{x}_c$ being the average score of the whole test; and $p_i$ being the difficulty of the i-th question item in relation to the multiple-choice test; is the standard deviation of the whole test score and is determined by the formula:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \qquad (3)$$

With $x_i$ being the score of the i-th candidate taking the test, $\bar{x}$ being the average score of the whole test, n is the number of examinees taking the test.

The correlation coefficient of objective question items ranges from -1 to 1. When the candidates who do the question correctly get a high score (the question has many correct answers) and at the same time the total score for the whole test of this candidate is also high, the correlation coefficient of the questions is close to 1. The correlation coefficient of the question item is close to -1 when the examinees who answer the question correctly have high scores but the scores of the test scores are low, and vice versa. The correlation coefficient of the question item is 0 if the score of the question item and the score of the whole test do not have a strong and stable relationship with each other (Lam Quang Thiep, 2011). Therefore, these question items need to be removed from the test.

***The reliability of the multiple choice exam papers***

The reliability of a test is a quantity showing the accuracy of the measurement executing on that test (Lam Quang Thiep, 2011). The reliability of a multiple-choice test is mainly influenced by three factors such as the correlation between the question items on the test, the length of the test, and the content of the whole test. Theoretically, the reliability of the test has a value from zero to one, corresponding to the confidence level from no confidence to very high confidence. One of the commonly used indicators to assess the reliability of the test is the Cronbach's Alpha value, with the expected value of the reliability coefficient reaching 0.8 or more (De Champlain, 2010; Downing, 2004). According to Brennan (2006), classroom tests have a good level of reliability when the reliability index is 0.7 or higher. Specifically, the value of 0.7 to 0.8 indicates that the reliability of the test is good, 0.8 to 0.9 corresponding to a very good level, and above 0.9 being considered as completely perfect.

**2.2. Research data**

The data used in this study are the results of the test papers on Basic Informatics, performed by the students majoring in Informatics Teacher education and Computer Science, belonging to university level of the formal training system in Dong Thap University from the academic year of 2017-2018 to that of 2020-2021. The number of question items and the number of students participating in each exam papers are shown in Table 1.

**Table 1. Description of the research data (the exam paper on Basic Informatics)**

| Academic year | Number of items | Number of test takers |
|---|---|---|
| 2017 - 2018 | 50 | 34 |
| 2018 - 2019 | 40 | 48 |
| 2019 - 2020 | 40 | 40 |
| 2020 - 2021 | 50 | 74 |

*Source: An extract of the authors' research data, 2022*

Basic Informatics is a course belonging to the general knowledge group, is compulsory for students of Informatics Teacher education and Computer Science and is organized in the first semester of the first year in the entire training program. Currently, in Dong Thap University, this course is instructed by many lecturers, so there are differences in the number of question items in the tests held over the years (2 exam papers of 40 questions each and 2 others of 50 question items each). Each question item in the exam paper is designed with 04 answer options, including 03 distractor options and 01 key option (correct answer). In addition, because the students enrolled in the majors of Computer Science and Informatics Teacher education vary in number from year to year, the number of students taking the exam varies from year to year, too. Specifically, the number of students taking the exam in the 2017-2018 academic year is the lowest, at about 34 students and in the 2020-2021 academic year the figure went to top, at about 74 students. As it is stipulated in the Regulations on Orgazining Exams by the University's office of Testing and Quality Assurance, questions for the final test paper will be designed by a single individual teacher or a group of teachers under the assigmment by the Head of department. After being designed, the questions will be sent to the Head of department for approval before being submitted to the Office of Quality Assurance. It means the questions are checked for their form, their accuracy of knowledge, appropriateness of the teaching contents prescribed in the teaching syllabus without for level of difficulty or discrimination.

**2.3. Data analysis**

Currently, many specialized software packages have the function of analyzing objective test items and have been applied in many studies (Bui Anh Kiet and Bui Nguyen Phuong, 2018; Nguyen Van Canh and Nguyen Quoc Tuan, 2020; Nguyen Phuoc Hai, 2017; Nguyen Thi Hong Minh and Nguyen Duc Thien, 2006; Nguyen Bao Hoang Thanh, 2008; Bui Ngoc Quang, 2017). However, estimating the parameters of objective question items based on exam papers by using specialized software may cause certain difficulties for some lecturers in terms of their familiarization and application. Within the scope of this article, the parameters of objective question items are calculated by Microsoft Excel software based on the definition and characteristics of each parameter. In addition, the evaluation of the question items is based on the parameters related to these question items. Specifically, a question item is considered satisfactory when the difficulty value reaches from 0.25 to 0.75, the discrimination value reaches from 0.2 or higher, and correlation coefficient of item is positive. In addition, the distractor options of the question items must appeal examinees to choose and there should be no big difference between the given options, and at the same time, the distractor options must have a negative correlation coefficient value.

**3. Findings and discussions**

**The reliability of the multiple choice exam papers**

The evaluation on reliability of the exam papers is done basing on Cronbach's Alpha value. The results of calculating the reliability coefficient Cronbach's Alpha from the data of basic Informatics exams over the school years are shown in Table 2.

**Table 2. Values of Cronbach's Alpha**

| No | Question tests on Basic Informatics | Cronbach's Alpha |
|---|---|---|
| 1 | Academic year 2017 - 2018 | 0.707 |
| 2 | Academic year 2018 - 2019 | 0.685 |
| 3 | Academic year 2019 - 2020 | 0.772 |
| 4 | Academic year 2020 - 2021 | 0.828 |

*Source: An extract of the authors' research data, 2022*

Statistics results show that the test question items show Cronbach's Alpha coefficient values of 0.685 or higher. Thus, among the tests under our investigation, 03 of them meet the requirements of reliability index when the value reaches 0.7 or higher and 01 test has an unsatisfactory value, with Cronbach's Alpha value being lower than 0.7. In addition to the reliability coefficient value, in this study, the authors focus on analyzing and evaluating the quality of each question item in each test based on the characteristic parameters of each individual item including the level of difficulty, discrimination, correlation coefficient and quality of the distractors. In a particular manner, the study will detail the data on the parameters of one exam paper (academic year 2020-2021) as an illustration before giving the evaluation results for the remaining exam papers. Based on the values of the parameters in Appendix 1, the results of the analysis and evaluation of the question items in the basic Informatics exams are shown as follows:

***Difficulty and discrimination level of the question items***

Of the 50 question items in the Basic Informatics exam paper in the 2020-2021 academic year, 30 question items met the requirements on difficulty level (accounting for 60%) and 43 question items met the requirements on the level of discrimination (accounted for 86%). In particular, 27 question items in the exam met the requirements of both difficulty level and discrimination level (accounting for 54%). Meanwhile, 20 question items were unsatisfactory in terms of difficulty level (accounting for 40%), namely items 2, 3, 7, 9, 10, 11, 13, 18, 19, 22, 24, 25 , 26, 27, 28, 30, 33, 36, 41, 44. Among them, 19 question items are very easy (over 75% of students answered correctly) and 01 question item is very difficult (less than 25% of students answered correctly) ). In addition, 07 question item in this exam did not meet the requirements of discrimination, accounting for 14%, including items 4, 9, 11, 24, 33, 42, 49. Therefore, the exam paper contains 27 question items that are satisfactory in both difficulty and discrimination, accounting for 54%, while 04 question items are unsatisfactory in both difficulty and discrimination, accounting for 8%. By doing the same analysis, the evaluation results on the level of difficulty and discrimination in the rest exam papers are shown in Table 3.

**Table 3. Statistics on the level of difficulty for the exam papers**

| Academic year | **Very difficult** (P < 0.25) | | **Average** (0.25 ≤ P ≤ 0.75) | | **Very easy** (P > 0.75) | | **Total** |
|---|---|---|---|---|---|---|---|
| | Frequency | % | Frequency | % | Frequency | % | |
| 2017-2018 | 4 | 8.0 | 36 | 72.0 | 10 | 20.0 | 50 |
| 2018-2019 | 1 | 2.5 | 18 | 45.0 | 21 | 53.0 | 40 |
| 2019-2020 | 3 | 7.5 | 15 | 37.5 | 22 | 55.0 | 40 |
| 2020-2021 | 1 | 2.0 | 30 | 60.0 | 19 | 38.0 | 50 |
| **Total** | **9** | **5.0** | **99** | **55.0** | **72** | **40.0** | **180** |

*Source: An extract of the authors' research data, 2022*

The statistics in Table 3 show that the question items with an acceptable difficulty level (25% to 75% of students giving correct answers) range from 37.5% to 72%. Among them, the exam paper with the highest percentage of students who gave the correct answers is the one used in the 2017-2018 academic year and the lowest percentage of similar situation is the 2019-2020 academic year. In addition, most of the question items in the exam papers that do not meet the requirements of the difficulty level are very easy (over 75% of students giving correct answers), with a rate from 20% to 55%. Meanwhile, the exam papers contain very difficult questions (less than 25% of students answered correctly) but with a low rate, from 2% to 8%.

**Table 4. Statistics on the level of discrimination of the question items**

| **Academic year** | | | **2017-2018** | **2018-2019** | **2019-2020** | **2020-2021** | **Total** |
|---|---|---|---|---|---|---|---|
| Poor | (D ≤ 0.19) | Count | 19 | 15 | 13 | 7 | 54 |
| | | % | 38 | 37.5 | 32.5 | 14 | 30 |
| Acceptable | (0.2 ≤ D ≤ 0.29) | Count | 7 | 6 | 8 | 9 | 30 |
| | | % | 14 | 15 | 20 | 18 | 16.7 |
| Fairly good | (0.3 ≤ D ≤ 0.39) | Count | 11 | 10 | 4 | 9 | 34 |
| | | % | 22 | 25 | 10 | 18 | 18.9 |
| Excellent | (D ≥ 0.4) | Count | 13 | 9 | 15 | 25 | 62 |
| | | % | 26 | 22.5 | 37.5 | 50 | 34.4 |
| **Total** | | | **50** | **40** | **40** | **50** | **180** |

*Source: An extract of the authors' research data, 2022*

The statistics in Table 4 show that most of the question items used in the exam papers have a discrimination level of intermediate or higher (acceptable level). Particularly, the highest percentage of discrimination level is the exam paper for the 2020-2021 academic year, at 86% and the lowest percentage discrimination level is the exam paper for the 2017-2018 academic year, at 62%. However,

statistics show that in the exam papers, there remain many unsatisfactory question items in term of level of discrimination, especially the ones for the 2017-2018 academic year, at 38%, in the 2018-2019 academic year, at 37.5% and in 2019-2020 academic year, at 32.5% respectively.

In addition, the quality of objective test items is greatly influenced by the quality of the distractor options. The results of the evaluation on the distractor options of the question items in this exam paper are shown through the following analysis.

***Distractor options of obvious recognition and poor appeal to students' attention***

The statistics in Appendix 1 show that 13 question items in this test contain obvious distractor options, so they fail to appeal students. Specifically, the question item include: Question item 2 (option A), Question item 4 (option A), Question item 8 (option D), Question item 10 (option D), Question item 11 (option A, B), Question item 12 (option D), Question item 13 (option B), Question item 17 (option B), Question item 18 (option B, D), Question item 22 (option C), Question item 24 (option A, B ), Question item 28 (option D) and Question item 37 (option A).

***Distractor options of positive correlation coefficient***

Another aspect that can help show the quality of the distractor options in the question item is the value of the correlation coefficient of that option in comparison with the question item itself. Specifically, a distractor option with a positive correlation coefficient value is considered poor quality, because it appeals more high-qualified examinees than low-qualified candidates. This is unreasonable for distractor options to appear in an objective test. Statistical results in Appendix 1 show that this test contains 10 question items with distractor options of positive correlation coefficients, namely: Question item 5 (option C), Question item 6 (option A), Question item 9 ( option D), Question item 16 (option C), Question item 25 (option B), Question item 32 (option C), Question item 33 (option C), Question item 39 (option C), Question item 42 (option A and D) and Question 49 (option D). It is noticeable that the 03 question items in this exam contain the obvious false options and with positive correlation coefficient value, namely Question item 4 (option C), Question item 12 (option C). and Question item 24 (option C). In addition, the statistical results in Table 6 show that some question items contain options that cannot estimate the value of the correlation coefficient (represented by the symbol *). These options do not appeal examinees for their choice, so the correlation coefficient cannot be calculated.

The analysis results of the options in each question item showed that most of the unsatisfactory question items fail to meet the requirements of difficulty and discrimination level contained unsatisfactory distractor options. However, several question items not only are satisfactory both in terms of difficulty and discrimination level but also contain poor quality distractor option, thus need further consideration for adjustment and improvement. This shows that improving the quality of options in objective test items will help increase the quality of those question items. This is one of the important clues for the test designers to promptly detect unsatisfactory question items and take action to adjust and improve the weak items, contributing to improving the quality of the question items, thereby giving the accurate assessment to the learners' ability. By the above-mentioned data analysis and evaluation, we can show the amount of question items with unsatisfactory distractor options, and they are shown in Table 5.

The analysis results of the distractor options in the Basic Informatics exam papers show that the number of question items containing unsatisfactory distractor options in the exam paper is quite high, from 46% to 70%. This has greatly affected the difficulty level, the discriminatory level of the question item, generating question items of poor quality, which are not meaningful in accurately measuring learners' ability. The detection of low-quality distractor options is very important because it helps test designers use scientific ground to adjust and improve the quality of question items so as to enable accurate and effective assessment of the students' ability.

By applying Classical Test Theory and applying data analysis tools of Microsoft Excel to calculate the parameters of question item in objective tests, the study has shown how to analyze, evaluate objective question items in order to detect question items of good quality as well as point out unsatisfactory ones that need to be adjusted and improved. By analysing the collected data about 04 Basic Informatics exam papers used in Dong Thap University from the 2017-2018 to the 2020-2021 academic year, the study shows that these test papers contain quite a few

**Table 5. Statistics of question items with unsatisfactory distractor options**

| Exam paper (academic year) | | 2017 - 2018 | 2018 - 2019 | 2019 - 2020 | 2020- 2021 |
|---|---|---|---|---|---|
| Number of items | | 50 | 40 | 40 | 50 |
| Unsatisfactory distractor options | No attraction to students | 13 items (26%) | 19 items (47.5%) | 18 items (45%) | 13 items (26%) |
| | Negative correlation coefficient | 22 items (44%) | 9 items (22.5%) | 9 items (22.5%) | 10 items (20%) |
| | No attraction to students and negative correlation coefficient | 2 items (4%) | 6 items (15%) | 5 items (12.5%) | 3 items (6%) |
| | **Total** | **35 items (70%)** | **28 items (70%)** | **27 items (67.5%)** | **23 items (46%)** |

*Source: An extract of the authors' research data, 2022*

question items of unsatisfactory values, which need adjusting, improving or may be eliminated from exam papers due to serious violations of the parameters. In particular, the most common problem in the exam papers is that the appearance of easy question items accounts for a high rate, to 55%, followed by the question items of poor discrimination level with the highest rate of 38%. In addition, the quality of the distractor options is also a matter of concern because it greatly affects the quality of the question items. The analysis results show that the number of question items containing unsatisfactory distractor options in the exam papers accounts for a very high rate, from 46% to 70%. The appearance of many unsatisfactory question items in the above exam questions comes from many reasons, of which the most basic one is the unscientific work of question design. To be specific, the question items have not been tested and analyzed and evaluated before being introduced into official use. Therefore, unsatisfactory question items are not detected in time for adjustment and improvement, leading to a decrease in the quality of the exam papers. In addition, an exam paper with many low-quality question items will reduce the objectivity of the assessment results, especially not accurately assessing learners' abilities.

**4. Conclusions**

Based on the results of this study, the authors believe that in order to effectively apply the objective test to assess learning outcomes, question items used in exam papers must be tested, analyzed and evaluated before being officially used in the exams. In addition, if any courses whose tool of assessment is objective tests are required, then a bank of questions is needed, in which the question items in the question bank need to be scientifically designed, with strong verification by experts, especially via analyzing and evaluating the quality of each question item. On that basis, the test designers will choose the question items of good quality and promptly detect the bad ones for adjustment, improvement, or removal (in case the violation is serious).

**References**

Airasian, P. W. (1994). *Classroom assessment*. New York: McGraw-Hill.

Al-Wardy, N. M. (2010). Assessment methods in undergraduate medical education. *Sultan Qaboos University Medical Journal*, *10*(2), 203-209.

Ary, D., Jacobs, L. C., and Razavieh, A. (2002). *Introduction to research in education*. California: Wadsworth Group.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Baker, F. (2001). The basics of Item Response Theory. ERIC: *Clearinghouse on Assessment and Evaluation*, University of Maryland, College Park, MD.

Brennan, L. R. (2006). *Educational measurement* (4th ed.,). Washington DC: American Council on Education.

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

Bui Ngoc Quang. (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.

Carroll, R. G. (1993). Evaluation of vignette-type examination items for testing medical physiology. *Advances in Physiology Education*, *264*(6), 11-15.

Considine, J., Botti, M., and Thomas, S. (2005). Design, format, validity, and reliability of multiple - choice questions for use in nursing research and education. *Collegian*, *12*(1), 19-24.

Cox, K. R. (1976). How did you guess? Or what do multiple choice questions measure? *Medical Journal of Australia*, (1), 884–886.

Cronbach, L. J., and Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, *64*(3), 391-418.

Davies, A. (1990). *Principles of language testing*. Oxford: Blackwell.

De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education*, *44* (1), 109-117.

Downing, S. M. (2004). Reliability: On the reproducibility of assessment data. *Medical Education*, *38*(9), 1006-1012.

Duong Thieu Tong. (2005). *Testing and measurement academic achievement*. Hanoi: Social Sciences Publishing House.

Ebel, R. L. (1972). *Essentials of educational measurement*. New Jersey: Prentice Hall.

Fowell, S. L., Southgate, L. J., and Bligh, J. G. (1999). Evaluating assessment: the missing link? *Medical Education*, *33*(4), 276-281.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. New York: Routledge.

Hambleton, R. K., Swaminathan, H., and Rogers, J. H. (1991). *Fundamentals of Item Response Theory*. Newbury Park: Sage.

Johnstone, A. H., and Ambusaidi, A. (2000). Fixed response: what are we testing?. *Chemistry Education*: Research and Practice in Europe, *1*(3), 323-328.

Kheyami, D., Jaradat, A., Al-Shibani, T., and Ali, F. A. (2018). Item analysis of multiple-choice questions at the department of paediatrics, Arabian Gulf University, Manama, Bahrain. *Sultan Qaboos University Medical Journal*, *18*(1), 68-74.

Lam Quang Thiep. (2011). *Measurement in education - theory and application*. Hanoi: Vietnam National University Publishing House.

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

Nguyen Phuoc Hai. (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 Thi Hong Minh and Nguyen Duc Thien. (2006). Measurement assessment in the objective test: Question difficulty and examinees' ability. *Vietnam National University Journal of Science, (4)*, 34-47.

Nguyen Van Canh and Nguyen Quoc Tuan. (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.

Nicol, D. (2007). E-assessment by design: using multiple-choice tests to good effect. *Journal of Further and Higher Education, 31*(1), 53-64.

Odukoya, J. A., Adekeye, O., Igbinoba, A. O., and Afolabi, A. (2018). Item analysis of university-wide multiple choice objective examinations: the experience of a Nigerian private university. *Quality & Quantity, 52*(3), 983-997.

Patil, P. S., Dhobale, M. R., and Mudiraj, N. R. (2016). Item analysis of MCQs'-Myths and realities when applying them as an assessment tool for medical students. *International Journal of Current Research and Review, 8*(13), 12-16.

Scouller, K. (1998). The influence of assessment method on students' learning approaches: multiple choice question examination versus assignment essay. *Higher Education (35)*, 453-472.

Zubairi, A. M., and Kassim, N. A. (2006). Classical and Rasch analyses of dichotomously scored reading comprehension test items. *Malaysian Journal of ELT Research, 2*(1), 1-20.

**Appendix 1. Parameters of question items in the exam papers of Basic Informatics in 2020 - 2021**

| Item | Key | D | P | % answer | | | | Correlation coefficient | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | A | B | C | D | A | B | C | D |
| Q1 | C | 0.41 | 0.58 | 16.2 | 12.2 | **58.1** | 13.5 | -0.22 | -0.11 | 0.40 | -0.24 |
| Q2 | D | 0.25 | 0.84 | 0.0 | 13.5 | 2.7 | **83.8** | * | -0.14 | -0.14 | 0.19 |
| Q3 | B | 0.41 | 0.87 | 4.1 | **86.5** | 2.7 | 6.8 | -0.42 | **0.37** | -0.07 | -0.12 |
| Q4 | D | 0.06 | 0.72 | 0.0 | 8.1 | 20.3 | **71.6** | * | -0.11 | 0.04 | **0.03** |
| Q5 | B | 0.43 | 0.53 | 21.6 | **52.7** | 12.2 | 13.5 | -0.12 | 0.41 | 0.02 | -0.47 |
| Q6 | B | 0.34 | 0.64 | 23.0 | 63.5 | 12.2 | 1.4 | 0.02 | 0.33 | -0.44 | -0.20 |
| Q7 | C | 0.47 | 0.85 | 4.1 | 2.7 | **85.1** | 8.1 | -0.33 | -0.10 | **0.52** | -0.38 |
| Q8 | A | 0.42 | 0.74 | **74.3** | 12.2 | 13.5 | 0.0 | **0.39** | -0.45 | -0.06 | * |
| Q9 | C | 0.07 | 0.93 | 2.7 | 2.7 | **93.2** | 1.4 | -0.24 | -0.10 | **0.21** | 0.04 |
| Q10 | A | 0.47 | 0.76 | **75.7** | 4.1 | 20.3 | 0.0 | **0.35** | -0.23 | -0.26 | * |
| Q11 | D | 0.01 | 0.96 | 0.0 | 0.0 | 4.1 | **95.9** | * | * | -0.10 | **0.10** |
| Q12 | A | 0.56 | 0.64 | **63.5** | 31.1 | 5.4 | 0.0 | **0.43** | -0.46 | 0.01 | * |
| Q13 | C | 0.48 | 0.77 | 16.2 | 0.0 | 77.0 | 6.8 | -0.32 | * | **0.34** | -0.09 |
| Q14 | B | 0.45 | 0.68 | 20.3 | **67.6** | 9.5 | 2.7 | -0.21 | **0.42** | -0.28 | -0.20 |
| Q15 | D | 0.41 | 0.73 | 1.4 | 16.2 | 9.5 | **73.0** | -0.01 | -0.21 | -0.25 | **0.35** |
| Q16 | A | 0.49 | 0.37 | **36.5** | 14.9 | 28.4 | 20.3 | **0.31** | -0.17 | 0.03 | -0.25 |
| Q17 | A | 0.32 | 0.49 | **48.6** | 0.0 | 39.2 | 12.2 | **0.27** | * | -0.22 | -0.08 |
| Q18 | C | 0.59 | 0.81 | 18.9 | 0.0 | **81.1** | 0.0 | -0.60 | * | **0.60** | * |
| Q19 | C | 0.65 | 0.77 | 2.7 | 6.8 | **77.0** | 13.5 | -0.19 | -0.28 | **0.57** | -0.41 |
| Q20 | D | 0.29 | 0.57 | 4.1 | 35.1 | 4.1 | **56.8** | 0.05 | -0.17 | -0.10 | **0.19** |
| Q21 | A | 0.55 | 0.41 | **40.5** | 14.9 | 5.4 | 39.2 | **0.47** | -0.18 | -0.14 | -0.27 |
| Q22 | A | 0.36 | 0.80 | **79.7** | 12.2 | 0.0 | 8.1 | **0.36** | -0.07 | * | -0.45 |
| Q23 | B | 0.52 | 0.53 | 5.4 | **52.7** | 40.5 | 1.4 | -0.12 | **0.41** | -0.28 | -0.33 |
| Q24 | D | 0.00 | 0.96 | 0.0 | 0.0 | 4.1 | **95.9** | * | * | 0.01 | -0.01 |
| Q25 | C | 0.43 | 0.78 | 8.1 | 2.7 | **78.4** | 10.8 | -0.45 | 0.10 | **0.38** | -0.16 |
| Q26 | A | 0.53 | 0.82 | **82.4** | 1.4 | 5.4 | 10.8 | 0.57 | -0.26 | -0.21 | -0.45 |
| Q27 | B | 0.24 | 0.86 | 9.5 | **86.5** | 2.7 | 1.4 | -0.09 | **0.31** | -0.26 | -0.33 |
| Q28 | A | 0.35 | 0.87 | **86.5** | 4.1 | 9.5 | 0.0 | **0.30** | -0.30 | -0.15 | * |
| Q29 | C | 0.49 | 0.74 | 5.4 | 4.1 | **74.3** | 16.2 | -0.32 | -0.03 | **0.40** | -0.26 |
| Q30 | C | 0.24 | 0.89 | 2.7 | 1.4 | **89.2** | 6.8 | -0.17 | -0.06 | **0.32** | -0.25 |
| Q31 | C | 0.40 | 0.60 | 14.9 | 1.4 | **59.5** | 24.3 | -0.14 | -0.20 | **0.30** | -0.17 |
| Q32 | D | 0.28 | 0.49 | 6.8 | 10.8 | 33.8 | **48.6** | -0.38 | -0.21 | 0.15 | **0.18** |
| Q33 | D | -0.04 | 0.14 | 1.4 | 21.6 | **63.5** | 13.5 | -0.15 | -0.19 | 0.26 | -0.09 |
| Q34 | D | 0.29 | 0.61 | 4.1 | 1.4 | 33.8 | **60.8** | -0.23 | -0.15 | -0.16 | **0.28** |
| Q35 | A | 0.31 | 0.43 | 43.2 | 24.3 | 5.4 | 27.0 | **0.22** | -0.25 | 0.02 | -0.01 |
| Q36 | A | 0.29 | 0.89 | 89.2 | 4.1 | 4.1 | 2.7 | **0.35** | -0.11 | -0.36 | -0.09 |
| Q37 | B | 0.45 | 0.69 | 0.0 | **68.9** | 18.9 | 12.2 | * | 0.37 | -0.34 | -0.12 |
| Q38 | D | 0.38 | 0.72 | 9.5 | 9.5 | 9.5 | **71.6** | -0.30 | -0.06 | -0.28 | 0.42 |
| Q39 | B | 0.34 | 0.35 | 9.5 | 35.1 | 25.7 | 29.7 | -0.31 | **0.31** | 0.08 | -0.20 |
| Q40 | B | 0.73 | 0.60 | 32.4 | **59.5** | 1.4 | 6.8 | -0.46 | 0.56 | -0.01 | -0.22 |
| Q41 | A | 0.42 | 0.87 | **86.5** | 9.5 | 2.7 | 1.4 | **0.47** | -0.34 | -0.15 | -0.33 |
| Q42 | B | 0.06 | 0.49 | 13.5 | **48.6** | 27.0 | 10.8 | **0.03** | 0.11 | -0.15 | 0.01 |
| Q43 | C | 0.27 | 0.66 | 13.5 | 8.1 | **66.2** | 12.2 | -0.10 | -0.11 | **0.30** | -0.24 |
| Q44 | C | 0.47 | 0.85 | 1.4 | 12.2 | **85.1** | 1.4 | -0.03 | -0.57 | **0.60** | -0.20 |
| Q45 | A | 0.48 | 0.69 | **68.9** | 16.2 | 12.2 | 2.7 | **0.41** | -0.21 | -0.30 | -0.10 |
| Q46 | C | 0.31 | 0.49 | 21.6 | 18.9 | **48.6** | 10.8 | -0.07 | -0.20 | 0.24 | -0.05 |
| Q47 | B | 0.54 | 0.69 | 9.5 | **68.9** | 6.8 | 14.9 | -0.08 | 0.46 | -0.29 | -0.32 |
| Q48 | A | 0.29 | 0.68 | **67.6** | 27.0 | 1.4 | 4.1 | **0.20** | -0.11 | -0.03 | -0.20 |
| Q49 | B | 0.03 | 0.37 | **44.6** | 36.5 | 5.4 | 13.5 | -0.01 | **0.09** | -0.37 | 0.13 |
| Q50 | A | 0.39 | 0.68 | **67.6** | 18.9 | 9.5 | 4.1 | **0.27** | -0.18 | 0.00 | -0.28 |

*Source: An extract of the authors' research data, 2022*