DOI 10.31489/2023Ped2/114-118

UDC 14.35.07

Zh.K. Tleshova*, Zh.A. Tusselbayeva, A.A. Ichshanova, A.I. Urazbekova, M.S. Zhenisbayeva

*Astana IT University, Kazakhstan*
*(*Corresponding author's e-mail: zhibek.tleshova@astanait.edu.kz)*

*ORCID ID: 0000-0001-5095-5436**

# The analysis of multiple-choice tests in a Professional English course at Astana IT University: difficulty, discrimination and distractor efficiency indexes

The objective of this study is to analyse the quality of multiple-choice questions by applying such evaluative tools as a difficulty index (DIF), a discrimination index (DI), and a distractor efficiency (DE). Our analysis was based on the results of quizzes conducted in Trimester III of 2020-2021 for the course of Professional English at Astana IT University. DIF, DI, and DE of midterm quiz results of first year students with different language levels were analysed using Microsoft Excel and Moodle LMS. This study will examine multiple-choice questions (MCQs) in terms of their effectiveness with recommendations for improvement.
The main three research questions are:
What is the difficulty index of midterm multiple-choice questions?
What is the discrimination index of midterm MCQ quizzes?
What is the distractor efficiency of midterm MCQ distractors?
As the study showed, the difficulty index of MCQs was below average as only 43 % (n=13) reached an acceptable level. Therefore, 57 % (n=17) of the questions that fall under the category of "too easy" should be revised in the test. The discrimination index evaluation revealed 100 % efficiency. As for the distractor efficiency, the study results equalled 88 %, which was a more than satisfactory level. By applying the DIF, DE and DI tools on a regular basis MCQ quizzes and distractors might be improved, leading to an annual accumulation of a better-quality pool of questions for a more effective student assessment.

*Keywords:* multiple-choice questions, distractor efficiency, discrimination index, difficulty index, Professional English, functional distractor, non-functional distractor, assessment.



## *Introduction*

Assessment is crucial in higher education since it ensures that the whole teaching and learning process is reviewed and improved. Variety of assessment tools are used to measure student achievement. According to Brown and Hudson (1998), these tools are categorized into three groups: selected-response assessments (true-false questions, matching questions and multiple-choice questions); constructed response assessments (fill-in the gaps, short-answer questions, performance assessments); and personal-response assessments also known as alternative assessments (portfolio and self- or peer assessments) [1].

In recent years, with more attention paid to alternative measurement tools in higher education, it has been observed that multiple-choice tests continue to be widely used as time-saving, covering large numbers of students, immediate results availability, etc. Therefore, there is a need to ensure the quality of multiple-choice questions.

The most popular type of selected-response assessment is the multiple-choice question test. With the switch to distance and blended learning modes MCQ format of testing has become even more common due to the obvious advantages in terms of development, conduction, and further evaluation/assessment. Aspiring validity and reliability in terms of covering major parts of content, McCoubrie (2004) states that multiple-choice question quizzes can be considered as highly effective assessment methods [2]. Popham (2005) has pointed out several advantages of MCQs such as versatility, easiness of handling student misunderstandings with particular items, the possibility of a high cognitive challenge to learners when developed effectively [3]. According to Douglas (2012), the assessment bias could be reduced by the use of MCQs as it provides a more objective measurement of test results compared to assessment tools [4]. Though testing is an effective assessment and instrument it regularly undergoes criticism.

MCQ quizzes should be proof-read and analyzed by various methods after they are developed and later upgraded or improved based on the student test results. One of these methods is the item scores analysis for each test item. In item score analysis, difficulty index (DIF), discrimination index (DI), and distractor efficiency (DE) values are calculated by D'Sa and Visbal-Dionaldo in 2017 [5]. The difficulty index is the ratio

of the number of correct answers given to an item by the total number of respondents. If most of the respondents answered correctly, this item is considered to be too easy. The discrimination index is the value that helps instructors distinguish high achiever students and low achiever students. To assess how non-functioning distractors in MCQs perform, the distractor efficiency is calculated through the analysis of all the options included in quizzes and the number of student option selections. Analyzing each item contributes to increasing the validity and reliability of the tests by revealing whether the questions work well or not. So, the study is needed to evaluate the questions and distractors.

In this study, we aim to evaluate the quality of MCQs within Professional English teaching in Astana IT University. The objectives of our descriptive analysis are to identify the quality of MCQs by calculating:

(1) the difficulty index (DIF);
(2) discriminating index (DI);
(3) the distractor efficiency (DE).

*Materials and methods*

The descriptive analytical study is conducted at Astana IT University among first year IT students with language levels B1-C1. The study includes 30 multiple-choice questions developed by university instructors of English language Program in 2020-2021 for Professional English course. During the midterm exam in Trimester 3, each student has taken Moodle based online quizzes with non-randomly selected 30 multiple-choice questions using client-based Safe Exam Browser software. MCQs comprise one stem and several options: 83 % (n=25) of MCQs have one key and three distractors, 10 % (n=3) have one key and 5 distractors, 7 % (n=2) one key and two distractors. The correct response is scored as one point, and the incorrect answer or unattempted item is scored as zero, with no penalty.

Microsoft Excel and Moodle have been used to analyze DIF, DI, and DE of midterm quiz results in Trimester III of Professional English course shared by four different language instructors. 126 answer scripts were retrieved from Moodle database system of four different instructors to analyze the difficulty index (DIF). DIF has been calculated based on the following formula, where H indicates the number of correct responses given by high achievers, L is the total number of correct answers given by low achievers, N is the total number of students in both high and low groups as indicated by Mahjabeen et al. (2017): DIF= [(H+L)/N] × 100 [6]. High and low achievers were chosen based on the "Quartile" concept determined by the internal criterion using student scores on the quiz being analyzed using Microsoft Excel as it was described by Ding and Beichner's study (2009) [7].

A total of 30 MCQs and 120 options (30 correct answers and 94 distractors) were analyzed individually on Moodle to identify distractor efficiency (DE). DE is calculated based on the choice frequency of incorrect options by students. There are two classifications of distractors: functional distractors (FD) and non-functional distractors (NFD). If a distractor is selected by more than 5 % of the total number of students taking the quiz, the distractor appears to be functional. D'Sa and Visbal-Dionaldo (2017) argue that NFD is chosen by less than 5 % [5].

*Results and discussion*

The application of formulas indicated in the materials and methods section of the article the following results are presented. Regarding the DIF of 126 students, a total 48 are categorized as high performers and low performers, 31 and 17 respectively. Quiz results of these 48 students are included for analysis. Remaining 78 student results are excluded from the study based on the conclusion that there might be a high chance of a correlation between high scores (100.0) and cheating due to technical issues with proctoring application the Safe Exam Browser. According to the analysis of the difficulty index for 30 MCQs, only 43 % (n=13) goes into an acceptable category. Among these high category MCQs, 5 questions fall under the category of having good DIF. The rest 57 % (n=17) MCQs appear to be too easy. No items were judged to be difficult (DIF<30) in the entire quiz. To group items in terms of difficulty index, the criteria of categorization in DIF has been applied: "too easy" = >70 %, "average" = b/w 30-70 %, "good" = b/w 50-60 %, "too difficult" = <30 % [6]. Mean ± SD and range of DIF, DI and DE have been illustrated in Figure 1.

| Parameters | Total% |
|---|---|
| Students (n) | 48 (high and low achievers) |
| MCQs (n) | 30 |
| Score Total | 30 |
| Score obtained<br>(Mean ± SD)<br>Range (%) | <br>70.9±16.2<br>26.67±100 |
| Difficulty index (%)<br>Mean ± SD<br>Range | <br>68.76±13.72<br>33-90 |
| Discrimination index<br>Mean ± SD<br>Range | <br>3.24±13.392<br>0.46-1.08 |
| Distractor efficiency (%)<br>Mean ± SD<br>Range | <br>65.32±25.78<br>0-100 |

Figure 1. DIF, DI, DE.

Concerning DE, Moodle based quiz comprised of total 30 MCQs and 94 distractors. Out of total 94 distractors, 88 % (n=83) are functional and only 12 % (n=11) are non-functional. 76.67 % (n=23) MCQs showed DE up to 100 % (n=11), 66.66 % (n=11), 60 % (n=1) respectively. The number of MCQs with distractor efficiency of 33.33 % is n=3 and 20 % is n=1. The number of questions with zero efficiency level is n=3. Number of distractors and categorization of MCQs according to distractor efficiency is shown in Figure 2 and Figure 3.

| Parameters | Total (n) |
|---|---|
| MCQs (Total) | 30 |
| Distractors (Total) | 94 |
| Functional Distractors | 83 |
| Non-Functional Distractors | 11 |
| MCQs with zero NFDs/ 4 FDs (DE=100 %) | 11 |
| MCQs with 1 NFDs / 3 FDs (DE=66.6 %) | 11 |
| MCQs with 2 NFDs / 3 FDs (DE=60 %) | 1 |
| MCQs with 2 NFD s / 2 FDs (DE=33.3 %) | 3 |
| MCQs with 4 NFDs / 1 FDs (DE=20 %) | 1 |
| MCQs with 3 or more NFDs / 1 or 0 FDs (DE=0 %) | 3 |

Figure 2. Number of distractors and categorization of MCQs according to distractor efficiency.

| | |
|---|---|
| Q1 — 66,66 % (C<5 %) | Q16 (6 options) 60 % |
| Q2 –3 options only 100 % | Q17 (6 options) 20 % |
| Q3 — 100 % | Q18 (6 options) 0 % |
| Q4 — 33,33 % (B, D<5 %) | Q19 — 66,66 % (A<5 %) |
| Q5 — 100 % | Q20 — 0 % |
| Q6 — 3 options only 100 % | Q21 — 100 % |
| Q7 — 66,6 % (C<5 %) | Q22 — 100 % |
| Q8 — 100 % | Q23 — 66,66 % |
| Q9 — 66,66 % (D<5 %) | Q24 — 100 % |
| Q10 — 66,66 % (C<5 %) | Q25 — 66,66 % (C<5 %) |
| Q11 — 66,66 % (B<5 %) | Q26 — 66,66 % (A<5 %) |
| Q12 — 100 % | Q27 — 100 % |
| Q13 — 0 % | Q28 — 100 % |
| Q14 — 33,33 % (B, C<5 %) | Q29 — 33,33 % |
| Q15 — 66,66 % (D<5 %) | Q30 — 66,66 % (D<5 %) |

Figure 3. DE analysis of each question by percentage

Discrimination index to differentiate low and high performer students is 100 %. This means that the test questions and options can enhance efficiency to categorize students' by their performance and for the instructors to reveal the question difficulty range.

Test results judged as easy resulted in high scores. This can be partly resolved if a test is assessed for difficulty index (DIF), discrimination index (DI) and distractor efficiency (DE). Such approaches provide test item writers feedback as to whether their items were acceptably discriminating and as hard or easy as they had supposed in the writing phase. Therefore, they will take greater efforts to make more challenging items in future tests. If students get high scores for a test, their ability is not necessarily advanced or highly proficient. Utilizing difficulty index (DIF), discrimination index (DI) and distractor efficiency (DE) potentially influences the quality of test writing and even course instruction. More challenging items can be assigned more weight and become more rewarding for students who answered them. The students who answer only easy questions will earn lower scores.

The difficulty index (DIF), discrimination index (DI) and distractor efficiency (DE) were used in this paper to assess the MCQ tests. The goal was to determine if these tools were applicable to language test assessment as they were previously applied in science, technology, engineering, and medical subjects. The evaluation took place in an IT university in which MCQ tests are used for online midterm testing in the course of English for Professional Purposes.

Ideally MCQ level of difficulty should be average with (30-70 %) with high DI (>0.25) and 100 % DE. 11, 14. In present research according to DIF criteria, out of total 30 MCQs, 43 % meet the criteria of an acceptable MCQ. Regarding the DI and DE, total 100 % and 88 % fall in the categorization of ideal MCQ respectively. In our study, mean and standard deviation for DIF, DI and DE were fallen in the category of good MCQs. As for the difficulty index, in our paper out of total 30 MCQs, 0 (0 %) MCQ was too difficult and 17 (57 %) were too easy. Total 8 and 5 MCQs (43 %) were in acceptable and good category respectively.

In our study, concerning the discrimination index, 30 (100 %) MCQs manifested high possibility to distinguish students gaining low and high marks.

The received data supports the conclusion that the difficulty index (DIF), discrimination index (DI) and distractor efficiency (DE) seem to be high for assessment of item difficulty and distractor efficiency.

However, issues were raised concerning 5 MCQs as to how to assess the efficiency of items with 3 distractors and 6 distractors. Nevertheless, there was sufficient data about 25 MCQs with 4 distractors.

The limitation of this study is that it took place in one university and with only 126 students and which is more the MCQ test was not randomized and with single correct answer, otherwise it would cause an extra difficulty in application of the mentioned test assessment tools. Thus, evaluation of questions through item analysis is sufficient approach to make the valid pool of MCQs.

## *Conclusion*

The difficulty index of the considered items illustrated insufficient difficulty as less than half (43 %) appeared to be acceptable. So, 57 % of items should be substituted in the test. Regarding the discrimination index and distractor efficiency the results demonstrated high efficiency 100 % and 88 % correspondingly. The systematic application of the DIF, DE and DI tools might improve MCQ items, distractors and, subsequently, the production of a good quality test bank for effective assessment of students. Revision of the tests by the above-mentioned instruments contributes to the growth of test development skills of instructors.

## References

1 Brown, J.D. & Hudson, T. (1998). The Alternatives in Language Assessment. *TESOL Quarterly*, *32*(4), 653-675. https://doi:10.2307/3587999

2 McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: A literature review. *Medical Teacher*, *26*(8), 709–712. https://doi.org/10.1080/01421590400013495

3 Popham, W.J. (2005). Test better, teach better: The instructional role of assessment. Association for Supervision and Curriculum Development.

4 Douglas, M., Wilson, J. & Ennis, S. (2012). Multiple-choice question tests: A convenient, flexible and effective learning tool? A case study. *Innovations in Education and Teaching International*, *49*(2), 111–121. https://doi.org/10.1080/14703297.2012.677596

5 D'Sa, J.L. & Visbal-Dionaldo, M.L. (2017). Analysis of multiple choice questions: Item difficulty, discrimination index and distractor efficiency. *International Journal of Nursing Education*, *9*(3), 109-137.

6 Mahjabeen, W., Alam, S., Hassan, U., Zafar, T., Butt, R., Konain, S. & Rizvi, M. (2017). Difficulty index, discrimination index and distractor efficiency in multiple choice questions. *Annals of PIMS-Shaheed Zulfiqar Ali Bhutto Medical University*, *13*(4), 310-315.

7 Ding, L. & Beichner, R. (2009). Approaches to data analysis of multiple-choice questions. *Physical Review Special Topics - Physics Education Research, 5*(2), 1-17. https://doi.org/10.1103/physrevstper.5.020103

Ж.К. Тілешова, Ж.А. Тусельбаева, А.А. Ищанова, А.И. Уразбекова, М.С. Женисбаева

## Astana IT University кәсіптік ағылшын тілі пәні бойынша бірнеше таңдауы бар тестті талдау: күрделілігі, дискриминация және дистрактор тиімділігінің көрсеткіштері

Зерттеудің мақсаты күрделілік көрсеткіші (DIF), дискриминация индексі (DI) және дистракторлардың тиімділігі (DE) сияқты бағалау құралдары арқылы бірнеше таңдау сұрақтарының сапасын талдау. Бұл сипаттамалық талдау Астана IT University «Кәсіби ағылшын тілі» курсы бойынша 2020-2021 оқу жылының III триместрінде өткен аралық тест нәтижелеріне негізделген. Ағылшын тілін меңгеру деңгейі әртүрлі бірінші курс студенттерінің аралық тестінің нәтижелері DIF, DE және DI әдістерімен және Microsoft Excel және Moodle LMS арқылы талданды. Бұл зерттеуде оның тиімділігін бағалау үшін бірнеше таңдау тестісі талданған және оны жетілдіру бойынша ұсыныстар берілген. Негізгі зерттеу сұрақтары мынадай: «Бірнеше таңдау тест сұрақтарының күрделілік көрсеткіші қандай?», «Бірнеше таңдау тесті үшін дискриминация индексі қандай?», «Қарастырылып отырған тесттің дистракторларының тиімділігі қандай?». Зерттеу көрсеткендей, сұрақтардың күрделілік көрсеткіші орташа деңгейден төмен күрделілікті анықтады, өйткені тек 43 %-ы (n=13) қолайлы болды. Тиісінше, «тым жеңіл» санатына жататын сұрақтардың 57 %-ы (n=17) тестте қайта қаралуы керек. Дискриминация индексін бағалау 100 % тиімділікті көрсетті. Дистракторлардың тиімділігіне келетін болсақ, зерттеу нәтижелері 88 %-ды құрады, бұл жоғары тиімділікті көрсетеді. DIF, DE және DI құралдарын жүйелі түрде қолдану студенттерді тиімдірек бағалау үшін бірнеше таңдау сынақтарын, дистракторларды және жоғары сапалы сұрақтар базасын әзірлеуді жақсарта алады.

*Кілт сөздер:* бірнеше таңдау сұрақтары, дистракторлардың тиімділігі, дискриминация индексі, күрделілік индексі, кәсіби ағылшын тілі, функционалдық дистрактор, функционалдық емес дистрактор, бағалау.

Ж.К. Тлешова, Ж.А. Тусельбаева, А.А. Ищанова, А.И. Уразбекова, М.С. Женисбаева

## Анализ тестов со множественным выбором по дисциплине профессионального английского языка в Astana IT University: показатели сложности, дискриминации и эффективности дистракторов

Целью исследования является анализ качества вопросов со множественным выбором с помощью таких оценочных инструментов, как показатель сложности (DIF), индекс дискриминации (DI) и эффективность дистракторов (DE). Данный описательный анализ основан на результатах промежуточного теста, который был в Astana IT University, на курсе «Профессиональный английский язык», триместр III, в 2020–2021 учебном году. Результаты промежуточного теста студентов первого курса с разным уровнем владения английским языком были проанализированы методами DIF, DE и DI и посредством Microsoft Excel и Moodle LMS. В настоящем исследовании проведен анализ теста множественного выбора для оценки его эффективности и приведены рекомендации по его улучшению. Основными исследовательскими вопросами являются: «Каков показатель сложности тестовых вопросов со множественным выбором?», «Каков индекс дискриминации теста множественного выбора?», «Какова эффективность дистракторов рассматриваемого теста?» Как видно из исследования, показатель сложности вопросов выявил сложность ниже среднего, так как только 43 % (n=13) были приемлемыми. Соответственно, 57 % (n=17) вопросов, подпадающих под категорию «слишком легкие», должны быть пересмотрены в тесте. Оценка индекса дискриминации показала 100 %-ную эффективность. Что касается эффективности дистракторов, то результаты исследования составили 88 %, что свидетельствует о высокой эффективности. Применение инструментов DIF, DE и DI на регулярной основе может улучшить тесты множественного выбора, дистракторы и разработку высококачественной базы вопросов для более эффективной оценки студентов.

*Ключевые слова:* тесты множественного выбора, эффективность дистракторов, индекс дискриминации, показатель сложности, профессиональный английский язык, функциональный дистрактор, нефункциональный дистрактор, оценивание.