DOI: https://doi.org/10.33193/JEAHS.20.2023.325
ISSN Online: 2709-071X
ISSN Print: 2709-0701
مجلة العلوم التربوية والإنسانية
Journal of Educational and Human Sciences
www.jeahs.com
Volume (20) February 2023
العدد (20) فبراير 2023

Analysis of MCQs in Summative Exam in English: Difficulty Index, Discrimination Index and Relationship between them

Dr. Rabab Mohammad Alareifi
Assistant Professor of Educational Technology, College of Education, University of Jeddah, Kingdom of Saudi Arabia
Email: ralareifi@uj.edu.sa

ABSTRACT

This paper aims to assess the quality of a summative test items to improve its ability to measure students' knowledge acquisition. This test was used in the English subject for11th grade students. This study was administered at a Western district secondary school in Saudi Arabia. The test consisted of 22 multiple-choice questions used to collect data from 94 students randomly. The Kuder-Richardson Formula 20 (KR-20) was used for test items, to determine the internal consistency reliability, which reaches a good reliability of $\alpha = 0.70$. Difficulty and discrimination indices were used as well to evaluate the quality of the test. In addition, the relationships between difficulty and discrimination indices are measured.

The difficulty index analysis showed that 50% of the items are in the average level, while the rest of the items fluctuate among too difficult, moderately difficult, and too easy levels. Moreover, the difficulty index analysis showed that 45.0% of the items are in the good level, while the other items ranged differently among poor, acceptable, and excellent levels. The Pearson correlation coefficient (r) to estimate the relationship between the difficulty index and the discrimination index has a value of (-0.936), which indicates that there is a statistically significant relationship at the level ($\alpha \leq 0.05$) between the difficulty index and discrimination index of the multiple-choice question summative test.

To enhance the quality of this test, to better assess students' knowledge acquisition, this study recommends that items with too difficult and too easy levels of difficulty index, and items with poor discrimination index are to be reviewed and modified by English experts. Moreover, reevaluation of the content validity by an English teacher could further improve its quality as well.

Keywords: Difficulty Index, Discrimination Index, Summative Test, Multiple-Choice Questions.

DOI: https://doi.org/10.33193/JEAHS.20.2023.325
ISSN Online: 2709-071X
ISSN Print: 2709-0701
مجلة العلوم التربوية والإنسانية
Journal of Educational and Human Sciences
www.jeahs.com
Volume (20) February 2023
العدد (20) فبراير 2023

Introduction

Assessment of student performance plays an imperative role in the instructional process. The main goal of assessment is to improve learning (Gronlund, 1998). Testing is one of the most popular types of assessments used in the educational field. The term achievement or performance tests refers to tests that are aimed to evaluate the knowledge, skills, and abilities attained by students in a field or in a subject area, in which they have received instruction (Frey, 2018). Multiple-choice questions (MCQs) are one type of questions mostly used in performance tests. This form of test has been used for decades and instructors depend mainly on them to evaluate students' comprehension and acquisition of knowledge. Developing a good quality of MCQs is not easy. It is challenging and time consuming.

Test construction usually follows a systematic method that includes two steps: developing test items and evaluating the test (Franzen, 2011). Developing test items means to construct it depending on the content and objectives of the subject (Franzen, 2011). Evaluating the test includes measuring its reliability and analyzing its items that refer to "a mixed group of statistics that are computed for each item on a test. The item analysis helps to determine the role of each item with respect to the entire test" (Boopathiraj & Chellamani, 2013, p.189). This analysis helps in modifying the test's items to improve its quality by editing or eliminating some items which may be used again in subsequent tests. It also helps the instructors to focus on content that needs more explanation or emphasis. Reliability, and Difficulty Index (DIF) and Discrimination Index (DI) are strategies used to evaluate the quality of the test. DIF and DI indices are the parameters used to evaluate the standard of MCQs in an examination where the standard of MCQs can be interpreted as excellent, acceptable, or poor (Pande, Pande, Parate, Nikam, & Agrekar, 2013).

Even though some instructors have used some form of tests' item analysis, there has been no previous attempt to use the same data to help in constructing other tests (Zubairi, 2006). More studies on evaluating and analyzing tests are needed to encourage instructors to evaluate tests before the administration process (Boopathiraj & Chellamani, 2013).

From this comes the aim of this study to answer the main research question: What is the quality of the MCQs summative test in English subject for the 11^{th} grade female students in a Western district's secondary school of Saudi Arabia?

From the main research question came the following research questions:

1. What is the DIF and DI levels of the MCQs English summative test for the 11^{th} grade female students?
2. Is there a statistically significant relationship at the level ($\alpha \leq 0.05$)

DOI: https://doi.org/10.33193/JEAHS.20.2023.325
ISSN Online: 2709-071X
ISSN Print: 2709-0701
مجلة العلوم التربوية والإنسانية
Journal of Educational and Human Sciences
www.jeahs.com
Volume (20) February 2023
العدد (20) فبراير 2023

between the DIF and DI of the MCQs English summative test for the 11th grade female students?

Purpose of the Study

The aim of this study is to evaluate the quality of test item's reliability, DIF and DI, and to determine whether there is any relationship between the DIF and DI of these items in a summative performance test in the English subject for the 11th grade female students at a secondary school in the Western district of Saudi Arabia.

Significance of the Study

This study is one of the few studies which helps instructors to evaluate the quality of test items to improve its ability to measure students' knowledge acquisition. Moreover, the results of this study can help researchers and instructors at other schools, who are planning to measure students' performance in the same subject, to use these MCQs and benefit from its statistics. This study will address the gap in the literature and respond to the need for more studies of tests' analysis to improve their quality. It might also work as guidance for instructors and instructional institutions to measure the quality of tests before administering them to the students in different educational levels. Additionally,, the results of this study provide information that might inspire instructors to change their way of teaching and offer more explanation for students in specific areas.

Methodology

Population and Sample

The population of this study refers to the 11th grade female students studying in one of the Western district secondary schools in Saudi Arabia. This population includes only students who are studying during semester 2 in the academic year of 2022-2023. Random sampling (*n= 94*) was adopted for this study.

Construction and Selection of Test Items

This test was developed by the English teacher to be used at one of the Western district's secondary schools in Saudi Arabia. She used the assessments included in the required textbook *Mega Goal 2.2 Student Book* by McGraw-Hill Education (2022) along with additional questions. The test covered the unit "There is no Place Like Home" of the required textbook. The purpose of this test was to measure students' acquisition of concepts and knowledge about ideas related to home, descriptions of things students are looking for in homes, words connected with directions for places, expressing requests, offers, promises, warnings and making decisions, and discussions of quotes and feelings about home.

Developing the test items included specifying the construct of interest, which assesses content knowledge for the unit "There is no Place Like Home", analyzing the learning content and learning objectives. Moreover, it included developing a test blueprint, a table of specification (TOS), which is.

> a tool used to ensure that a test or assessment measures the content and thinking skills that the test intends to measure. Thus, when used appropriately, it can provide response content and construct (i.e., response process) validity

DOI: https://doi.org/10.33193/JEAHS.20.2023.325
ISSN Online: 2709-071X
ISSN Print: 2709-0701
مجلة العلوم التربوية والإنسانية
Journal of Educational and Human Sciences
www.jeahs.com
Volume (20) February 2023
العدد (20) فبراير 2023

evidence. A TOS may be used for large-scale test construction, classroom-level assessments by teachers, and psychometric scale development. (Frey, 2018, p.1654-1655)

The teacher developed a modified TOS for the test, which is based on Bloom's (1965) Taxonomy. Depending on this TOS, the teachers decided to include 22 of MCQs with three possible answers in which the student must select one correct answer. After that, she began preparing a preliminary draft of the test that takes 45 minutes to be completed by students. The maximum score for this summative test was 100 points.

The teacher used different strategies for evaluating the test's validity. As TOS can provide validity evidence for test constructors (Frey, 2018), she also examined the content validity. Three English Language experts reviewed the test items and evaluated whether the test was a valid measure of the concepts being measured, the test items are directly related to the learning objectives, and the clarity of the items. The feedback was taken into consideration and the test's items were modified depending on these reviews.

Data Collection

In the second semester of 2022 the test administration was carried out with 94 female students in the 11th grade at a secondary school in Saudi Arabia's Western district. The examination was carried out after studying the whole unit of "There is No Place Like Home". Prior to starting the study, required permissions were obtained. Moreover, a consent form for participating in, and publishing the results of, this study was given to the students to be signed by their guardians before starting this study.

Test Reliability

For reliability, the Kuder-Richardson Formula 20 (KR-20) was used to evaluate the internal consistency of the MCQs. A high value of KR-20 reflects a strong relationship between a test's items, while a low value reflects a weaker relationship, where values range from 0 to 1 (Zimmerman, 1972). The results of this test showed that KR-20 reaches an acceptable level of reliability of $\alpha = 0.70$

Results

To answer the first research question: What is the DIF and DI levels of the MCQs English summative test for the 11th grade female students? the researcher calculated the DIF and the DI using Microsoft Excel. "Difficulty index (DIF), also called ease index, describes the percentage of students who correctly answered the item" (Hingorjo & Jaleel, 2012, p. 143). A higher value of DIF shows most of the students gave the correct answer, meaning the questions are easy to attempt. The range is from 0-100%. "Discrimination index (DI), also called point biserial correlation (PBC), describes the ability of an item to distinguish between high and

DOI: https://doi.org/10.33193/JEAHS.20.2023.325
ISSN Online: 2709-071X
ISSN Print: 2709-0701
مجلة العلوم التربوية والإنسانية
Journal of Educational and Human Sciences
www.jeahs.com
Volume (20) February 2023
العدد (20) فبراير 2023

low scores" (p. 143). Its range is 0-1. These distributions contribute to identifying questions that can be edited or removed if not performed well (Mahjabeen et.al, 2018). The items are listed according to the degree of DIF (too difficult, moderately difficult, average, too easy); and DI (poor, acceptable, good, excellent). The results of this question are shown in table and figures 1 to 4.

Table (1) DIF for Test Items

Question Number	Difficulty Index	Result
1	0.63	Average
2	0.79	Too easy
3	0.56	Average
4	0.54	Average
5	0.24	Moderately difficult
6	0.52	Average
7	0.53	Average
8	0.17	Too difficult
9	0.47	Average
10	0.43	Average
11	0.81	Too easy
12	0.29	Moderately difficult
13	0.47	Average
14	0.82	Too easy
15	0.49	Average
16	0.17	Too difficult
17	0.22	Moderately difficult
18	0.84	Too easy
19	0.47	Average
20	0.14	Too difficult
21	0.27	Moderately difficult
22	0.26	Moderately difficult

DOI: https://doi.org/10.33193/JEAHS.20.2023.325
ISSN Online: 2709-071X
ISSN Print: 2709-0701
مجلة العلوم التربوية والإنسانية
Journal of Educational and Human Sciences
www.jeahs.com
Volume (20) February 2023
العدد (20) فبراير 2023

Fig (1): Distribute the Items According to the DIF Categories

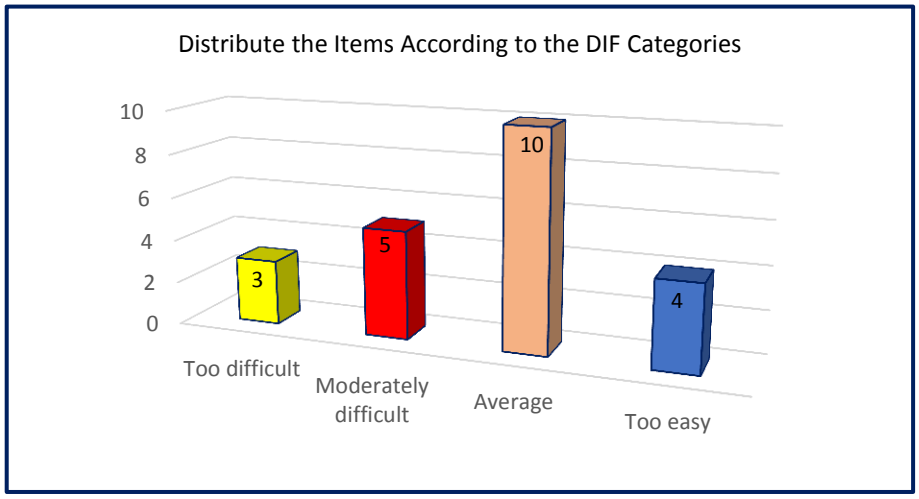

Table 1 and figure 1 show that the values of the DIF for all test items were between (0.14) and (0.84) ranging from too difficult to too easy. 3 items ranged between (0.14) and (0.17) which is a too difficult level, 5 items ranged between (0.22) and (0.29) which is a moderately difficult level, 10 items ranged between (0.43) and (0.63) which is an average level, and 4 items ranged between (0.79) and (0.84) which is a too easy level.

Table (2) Percentages for the Distribution of Test Items According to DIF Levels

N	Categories	Rang	Frequency	Percent
1	Too difficult	0.20 and Less	3	15.0%
2	Moderately difficult	0.21 to 0.30	5	25.0%
3	Average	0.31 to 0.70	10	50.0%
4	Too easy	0.71 and Above	4	20.0%

DOI: https://doi.org/10.33193/JEAHS.20.2023.325
ISSN Online: 2709-071X
ISSN Print: 2709-0701
مجلة العلوم التربوية والإنسانية
Journal of Educational and Human Sciences
www.jeahs.com
Volume (20) February 2023
العدد (20) فبراير 2023

Fig (2): Percentages for the Distribution of Test Items According to DIF Levels

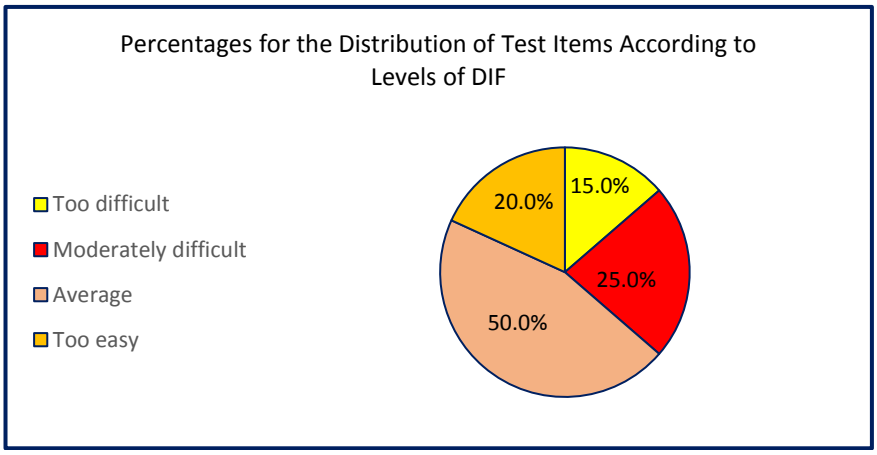

Table 2 and figure 2 show the percentages for the distribution of test items according to levels of DIF. It is cleared that 15.0% of the items came in the too difficult level, 25.0% of the items came in the moderately difficult level, 50.0% of the items came in the Average level, and 25.0% of the items came in the too easy level.

Table (3) DI for Test Items

Question Number	Discrimination Index	Result
1	0.21	Acceptable
2	0.13	**Poor**
3	0.25	Good
4	0.26	Good
5	0.72	Excellent
6	0.28	Good
7	0.27	Good
8	0.85	Excellent
9	0.33	Good
10	0.35	Good
11	0.15	**Poor**
12	0.65	Excellent
13	0.33	Good
14	0.17	**Poor**

DOI: https://doi.org/10.33193/JEAHS.20.2023.325
ISSN Online: 2709-071X
ISSN Print: 2709-0701
مجلة العلوم التربوية والإنسانية
Journal of Educational and Human Sciences
www.jeahs.com
Volume (20) February 2023
العدد (20) فبراير 2023

15	0.30	Good
16	0.85	Excellent
17	0.75	Excellent
18	0.19	**Poor**
19	0.33	Good
20	0.89	Excellent
21	0.68	Excellent
22	0.70	Excellent

Fig (3): Distribute the Items according to the DI Categories

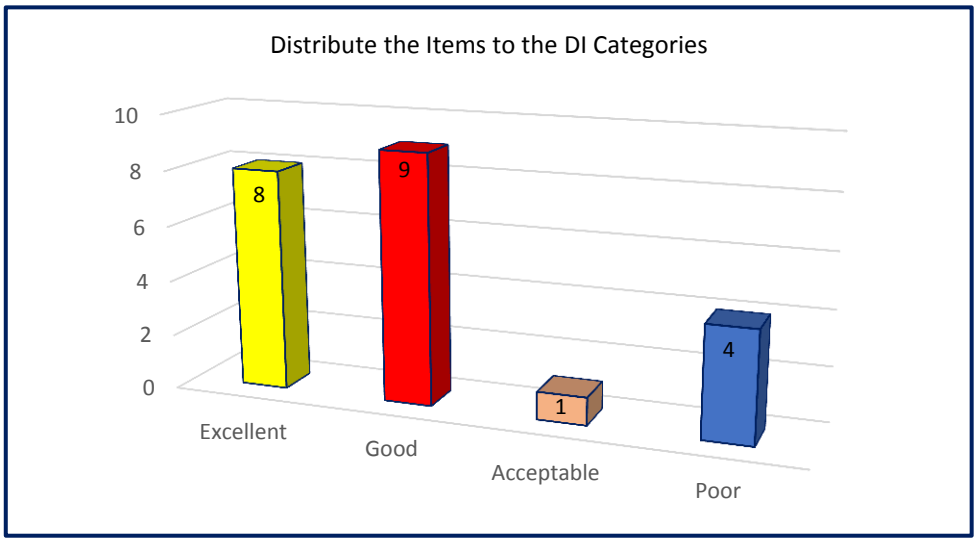

Table 3 and figure 3 show that the values of the DI for all test items were between (0.13) and (0.89) ranging from poor to excellent levels of DI. 8 items ranged between (0.65) and (0.89) which is in the excellent level, 9 items ranged between (0.25) and (0.35) which is in the good level, 1 item is located in (0.21) which is in the acceptable level, and 4 items ranged between (0.13) and (0.19) which is in the poor level.

DOI: https://doi.org/10.33193/JEAHS.20.2023.325
ISSN Online: 2709-071X
ISSN Print: 2709-0701
مجلة العلوم التربوية والإنسانية
Journal of Educational and Human Sciences
www.jeahs.com
Volume (20) February 2023
العدد (20) فبراير 2023

Table (4) Percentages for the Distribution of Test Items According to DI Levels

N	Categories	Rang	Frequency	Percent
1	Poor	0.20 and Less	4	20.0%
2	Acceptable	0.21 to 0.24	1	5.0%
3	Good	0.25 to 0.35	9	45.0%
4	Excellent	0.36 and Above	8	40.0%

Fig (4) Percentages for the Distribution of Test Items According to DI Levels

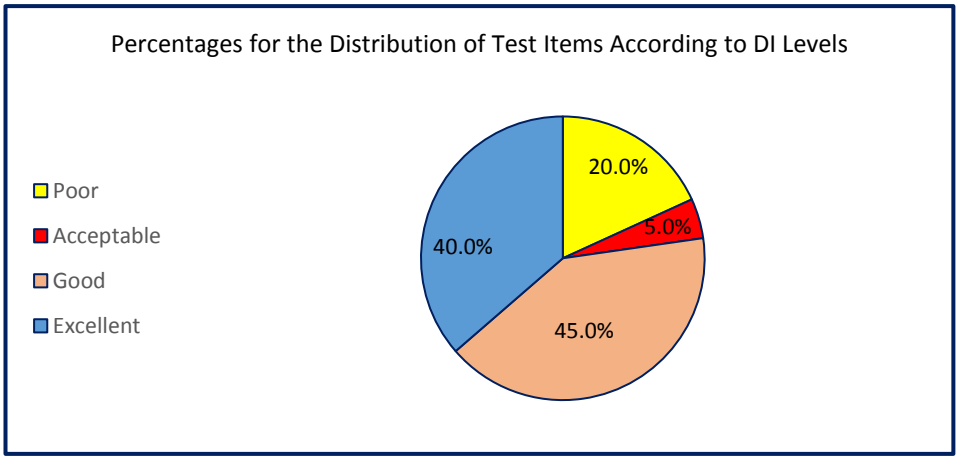

Table 4 and figure 4 show percentages for the distribution of test items according to levels of DI where 20.0% of the items came in the poor level, 5.0% of the items came in the acceptable level, 45.0% of the items came in the good level, and 40.0% of the items came in the excellent level.

To answer the second research question: Is there a statistically significant relationship at the level ($\alpha \leq 0.05$) between the DIF and DI of the MCQs English summative test for the 11th grade female students? the researcher used the Statistical Package for Social Sciences (SPSS) program to calculate the correlation via Pearson correlation coefficient (r) formula. The results of this question are shown in table 5 and figure 5.

DOI: https://doi.org/10.33193/JEAHS.20.2023.325
ISSN Online: 2709-071X
ISSN Print: 2709-0701
مجلة العلوم التربوية والإنسانية
Journal of Educational and Human Sciences
www.jeahs.com
Volume (20) February 2023
العدد (20) فبراير 2023

Table (5) The Relationship Between the DIF and DI of the MCQs Summative Test

Correlations		
		Test
Test	Pearson Correlation	-.936**
	Sig. (2-tailed)	.000
	N	22
**. Correlation is significant at the 0.01 level		

Fig (5) The Relationship Between the DIF and DI of the MCQs Summative Test

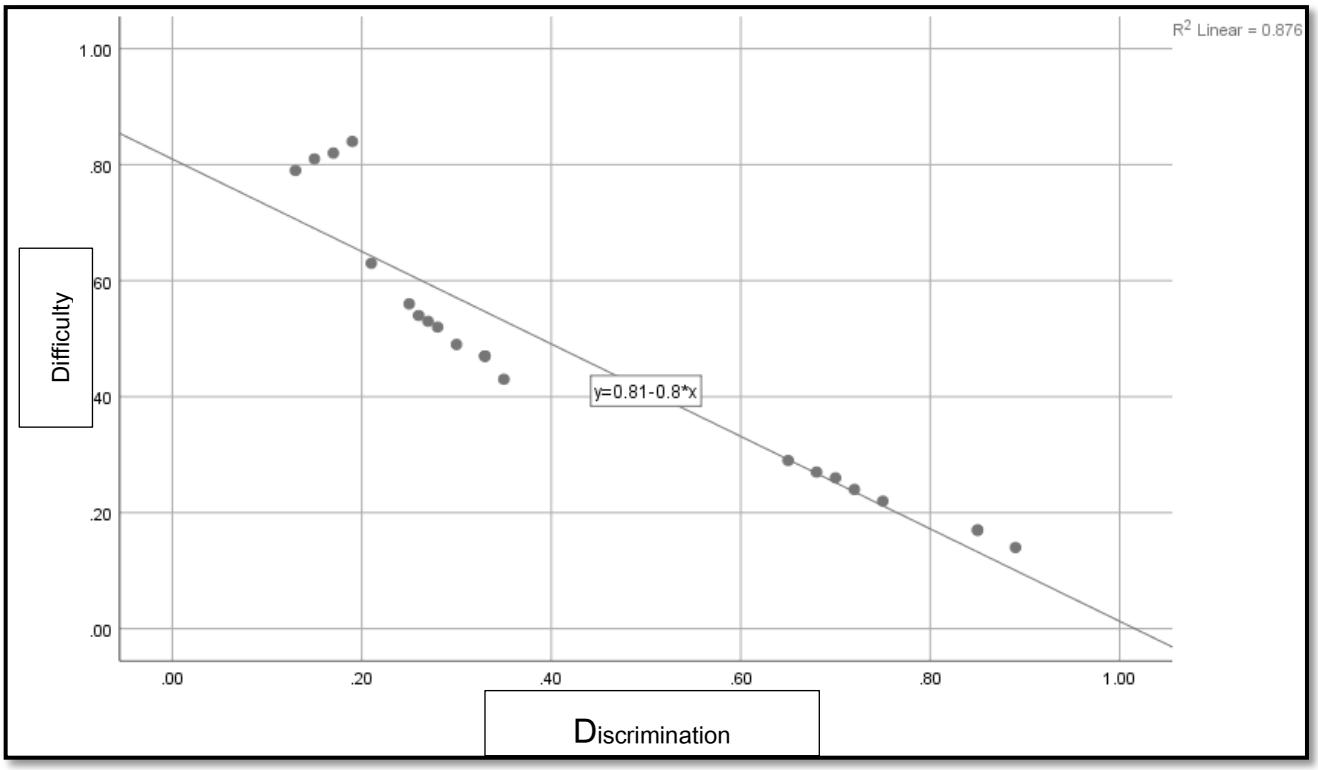

Table 5 and Fig 5 show that the value of Pearson correlation coefficient is (-0.936). This value of the Pearson correlation is a negative value, which indicates that it is an inverse relationship, which means the easier item has poor DI and the difficult item has excellent DI. This value indicates that there is a statistically significant relationship at the level ($\alpha \leq 0.05$) between the DIF and DI of the MCQs summative test.

DOI: https://doi.org/10.33193/JEAHS.20.2023.325
ISSN Online: 2709-071X
ISSN Print: 2709-0701
مجلة العلوم التربوية والإنسانية
Journal of Educational and Human Sciences
www.jeahs.com
Volume (20) February 2023
العدد (20) فبراير 2023

Discussion

For decades, performance tests have shown effective measurement of students' knowledge acquisition in different educational contexts. MCQs are one type of question that is mostly used in performance tests. Constructing a test should follow specific guidelines and its quality should be evaluated. Nevertheless, other factors might affect tests' results like timing, students' perceptions and individual differences. This paper evaluated the quality of MCQs items via measuring reliability, DIF and DI, and the relationship between DIF and DI.

The results showed that tests items reached an acceptable reliability level. Regarding the DIF, 50% of the items are in the average level, while other items fluctuate among too difficult, moderately difficult, and too easy levels. Regarding the DI, 45.0% of the items come in at a good level, while the rest of the items range differently among poor, acceptable, and excellent levels.

From this result, it is recommended that items with too difficult, and too easy of DIF, and items with poor DI be reviewed and modified by instructors and English experts. It is also recommended to readminister the MCQs test after modifying it and reevaluating its DIF and DI, which might further enhance the test in general. More reviews for the content validity by the English teacher could further improve its quality as well. Instructors are always recommended to pick test items that have average DIF and good DI to make sure that the test is efficiently measuring students' knowledge acquisition. More studies are needed to understand what factors other than validity, reliability, DIF and DI could improve the quality of tests in general, and MCQs in specific.

The results of this evaluation work as beneficial feedback to the future instructors who are planning to use it, about its quality and effectiveness. In addition, instructors would know which objective needs more explanation and emphasis when teaching the same subject in the future. It also might help in making decisions on whether this test is suitable or not for the level of students' English proficiency.

Conclusion

This paper assessed the quality of MCQs test items used in English subject for the 11th grade students as a summative assessment for their knowledge acquisition. Reliability, DIF and DI were measured and analyzed. Moreover, the relationship between DIF and DI was examined.

The test showed good reliability. Moreover, 50% of the test's items fulfilled the criteria of the average level of DIF. In addition, 45% of the items have a good level of DI, which means test items were able to discriminate the student's performance in the test. However, other items should be reviewed and modified to improve the test ability in assessing students' performance.

To sum up, this test can be used to assess students' performance levels, but it can be improved to provide better assessment. This improvement would help in

DOI: https://doi.org/10.33193/JEAHS.20.2023.325
ISSN Online: 2709-071X
ISSN Print: 2709-0701
مجلة العلوم التربوية والإنسانية
Journal of Educational and Human Sciences
www.jeahs.com
Volume (20) February 2023
العدد (20) فبراير 2023

constructing subsequent assessment tests. It is recommended to repeat the administration of the test items and reexamine its reliability, DIF and DI, and compare the results to this paper's findings.

References

1. Bloom, B. S. (1956). Taxonomy of educational objectives. Vol. 1: Cognitive domain. New York: McKay, 20-24.

2. Boopathiraj, C., & Chellamani, K. (2013). Analysis of test items on difficulty level and discrimination index in the test for research in education. *International journal of social science & interdisciplinary research, 2*(2), 189-193.

3. Frey, B. B. (Ed.). (2018). *The SAGE encyclopedia of educational research, measurement, and evaluation*. Sage Publications.

4. Gronlund, N. E. (1998). *Assessment of student achievement*. Allyn & Bacon Publishing, Longwood Division, 160 Gould Street, Needham Heights, MA 02194-2310; tele.

5. Hingorjo, M. R., & Jaleel, F. (2012). Analysis of one-best MCQs: the difficulty index, discrimination index and distractor efficiency. *JPMA-Journal of the Pakistan Medical Association, 62*(2), 142.

6. Mahjabeen, W., Alam, S., Hassan, U., Zafar, T., Butt, R., Konain, S., & Rizvi, M. (2018). 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. Pande, S. S., Pande, S. R., Parate, V. R., Nikam, A. P., & Agrekar, S. H. (2013). Correlation between difficulty and discrimination indices of MCQs in formative exam in physiology. *South-East Asian Journal of Medical Education, 7*(1), 45-50.

8. Zimmerman, D. W. (1972). Test reliability and the Kuder-Richardson formulas: Derivation from probability theory. *Educational and Psychological Measurement, 32*(4), 939-954.

9. Zubairi, A.M & Kassim. N.L.A. (2006) Classical and Rasch analysis of dichotomously scored reading comprehension test items. *Malaysian J of ELT Res*, 2, pp. 1-20.