British Journal of Education
Vol.4, No.8, pp.32-43, August 2016
Published by European Centre for Research Training and Development UK (www.eajournals.org)

A COMPARISON BETWEEN TWO TEST ITEM FORMATS: MULTIPLE-CHOICE ITEMS AND COMPLETION ITEMS

Dr. Ahmad M. Thawabieh

Educational Psychology Department, Faculty of Educational Sciences, Tafila Technical University, Jordan.

ABSTRACT: *This study aimed at investigating the influence of item formats upon students' performance and item psychometric characteristics. The students of evaluation and measurement 68 students were chosen to collect the data, 2 test format were applied; Multiple Choice items test and Completion items test. The results indicated that students performance was effected by the type of the test items, female students' performed higher than male students' on the 2 forms, and the psychometric characteristics of the test were found to be affected by test format.*

KEYWORDS: Test, Psychometric Characteristics, Performance, Item, Item Format.

INTRODUCTION

Choosing the test items for assessing the students' achievement is considered to be one of the greatest elements in test construction. Items differ in their degree of the freedom given to the students to express themselves and the skills and knowledge they acquired (Allam, 2007).

Flucher and Davidson (2007) indicated that items or tasks that first come to the mind during test construction. So test constructers must choose the best items to assess students' achievement. (Allam. 2007), Alzude and Alaan (2005), Kufahi (2003), and Thorndike and Hagen (1986) concentrated on the following rules for selecting the item format: item format must be suitable to the learning outcomes, students' age, item difficulty, appropriate to the assessment objectives, content and teachers' experience. Phipps and Brackbill (2009) indicated that assessment items should correspond to the content (lecture material and readings) learning objectives and suitability match the instructional methods. Crocker and Algina (1986) characterized the test developer's task as requiring two major types of decisions: what to measure and how to measure, according to that they have to concentrate upon the following when developing the pool of items: selecting an appropriate item format, and verifying the proposed format which is feasible for the intended examinees.

Multiple choice (MC) item is consisted of: a stem which provides the examinee with the problem that he/ she has to solve, the correct answer and the alternatives. MC items were widely used in testing, this is because of its advantages, like objectivity and high reliability. The test writer can measure variety of skills and learning outcomes, they are scored easily, rapidly, accurately and objectively by teachers, scoring machines or computers, the high coverage of the course objectives, and the last but not the least is their ease of administration. Although they may have items which have the following shortcomings: high guessing ratio, easy to be cheated, and it needs professional test writer, especially writing effective alternatives.

Harries (1969), McNamara (2000), Fuchler and Davidson (2007) and Haynie (1994) indicated that students' performance on MC items is superior to that on short answer questions.

ISSN 2055-0219(Print), ISSN 2055-0227(online)

Completion items format is a question in which students are asked to answer a short question or finish an incomplete statement by filling in a blank with the correct word, number, symbol, or short phrase. These items have the following advantages: wide content coverage, minimize guessing as compared to multiple choice or true-false item and can be scored objectively. However hey have the following limitations: it measures the lower levels of cognitive abilities, item consuming to score when compared to true – false and MC items, more difficult to score because some items may have more than one answer, they often include more irrelevant clues than do other item types, and not free of subjectivity (center of innovations).

After assessment, instructors must evaluate the items that used in evaluating their students, because the item characteristics may affect the students' achievement. The two main common item characteristics are: item difficulty and item discrimination. Item difficulty refers to the percentage of examinees who answered the item correctly, the values for item difficulty ranges from (0%-100%), items with difficulty below 30% were considered to be difficult, while items with difficulty higher than 70% were considered to be easy. If the item has a low difficulty (less than 30%) there are several possible causes: the item may have been miskeyed, the item challenging the level of the students ability, or the item may be ambiguous. But, if the item has a higher difficulty (more than 70%), this could be explained by: the item is too easy, the item may have been miskeyed or ineffective alternatives (Najar, 2010; Alam, 2007).

Item discrimination index is the second important characteristics of the item. It indicates the ability of the item to differentiate between the students who scored well and who scored poorly in the test. Item discrimination index ranges between -1 and +1, the positive values are desirable, items with negative and zero values should be reviewed, zero discrimination indicates that the item does not differentiate between students. Negative and low discrimination index may result by: miskeyed items or ambiguous items.

Many studies were conducted to explore the effect of item formats upon students achievement and test characteristics. The study of Culligan (2015) aimed to compare three common vocabulary test formats, the Yes/ No test, the vocabulary knowledge scale, and the vocabulary level test, as measures of vocabulary difficulty. The three tests were given to 165 Japanese students, the results indicated that the three tests measured one major latent trait (unidimensional) and they were significantly correlated in estimating their item difficulty.

The study of Simbak, Aung, Ismail, Joush, Ali, Yaseein, Haque, and Rebuan (2014) aimed to compare between students' performance on two evaluation techniques: multiple true-false and single best answer test formats, and correlated them with other assessment outcomes. The study analyzed the data for 20 item formats for each type of the questions, the participants were 3rd year medicine students at Sultan Zainal Abdin University in Malaysia. The results indicated that students got higher marks in single best answer than multiple true false quiz. Single best answer test results were found to be well correlated with clinical marks.

Hudson (2012), investigated the effectiveness of two forms of questions: MC and short answer questions used in the State University Entrance Examination for Chemistry and the effect of gender on performance, data was collected from 192 year 11 students from 4 secondary colleges. The researcher constructed short test that asked similar questions but in both types: multiple choice and short answer form. The results indicated that male students achieved higher scores than female students with respect to mean scores on both tests and subtests.

ISSN 2055-0219(Print), ISSN 2055-0227(online)

Many studies were conducted to indicate the effect of item format on students achievement. Phipps and Brackbil (2009) evaluated the relationship between assessment item formats (case–based and non-case– based) and item performance characteristics, they collected 1575 items from examinations administered in several therapeutics courses over 4 academic years. The results indicated that non case- based items demonstrated a higher discrimination index than case-based items, while case-based items were lengthier, included more detailed information and not more difficult.

Demas (1998) conducted a study to explore the effect of gender differences in math and science performance on the Michigan High School Proficiency Test, and the relationship between sex, item format and students' ability. The sample consisted of 102 schools that participated in math test(1392 female and 1290 male) and (99) other schools participated in science test (1352 female and 1294 male).The results indicated a small interaction in science and nonexistent interaction in math when students of all ability levels were considered when only the highest ability students were considered, male students scored higher on the multiple choice section, whereas female students either scored higher on the constructed response section or the degree to which the male students scored higher was less on the constructed response section.

Problem of Statement

The study emerged from the researcher observations of the students attitudes and behavior toward assessment process and item formats they prefer, their usual repeated question after scheduling the test time and content was: what is the item format of the test items? Are they essay items or MC items. Based on the previous result, the researcher conducted this study in order to answer the following questions:

1. Do the students' performance affected by item formats?
2. Are there gender differences in students' performance attributed to the item formats?
3. Do the psychometric characteristics vary by item formats (MC and completion items)?

METHODOLOGY

Design

Quasi-experimental design was used to collect data, 2 test format were applied for this purpose.

Sampling Procedure

The tests were administered to 68 2nd year students attending measurement and evaluation course at college of education at Tafila Technical University. In the fall semester 2015 they were divided randomly into 2 groups, each group consisted of 34 students from both sexes, group 1 subjected to MC test format, while, group 2 was subjected to completion format test.

Instruments

The instruments used to collect data were 2 measurement and evaluation test formats, form A composed of 22 MC items. Form B composed of 22 completion items, the 2 forms assessing the same content, the items in the 2 forms were identical in the objective they attempted to

ISSN 2055-0219(Print), ISSN 2055-0227(online)

assess, the only difference was the item format (appendix 1), validity of the tests were checked using content validity.

RESULTS

The study aimed to investigate the effect of item formats upon item characteristics and students' performance.

To answer the 1st question (Is the students' performance affected by item format?). descriptive statistics (means and standard deviations) were used. Table 1 represents the findings for this question

Table (1). Means and standard deviation for students' performance

	Multiple choice	Completion
Means	13.45	7.06
Standard deviations	2.88	28.52

According to table 2, It was found that the students' performance on multiple choice items test was higher (13.45) than Completion items test (7.06). To examine the effect of item format upon students' performance, table 2 represents the findings of that.

Table (2). Independent Samples T-test for students' performance

	t	df	Sig
Performance	6.197	64	.000

Table 2 represents the t-test for independent samples for students' performance, it shows that the difference in means was statistically significant (α=0.05) in fever of MC items test.

To answer the 2nd question (Are there gender differences in students' performance attributed to the item formats?). Means and standard deviations were used. Table 3 represents the students' performance according to their gender

Table (3). Descriptive statistics for gender performance

	Male		Female	
	mean	S. D	mean	S. D
MC	11.43	3.20	14.00	3.21
Completion	5.14	1.62	7.58	5.35

As shown in table 3 female students had higher performance on the 2 tests format. In order to investigate if the difference in means was significant; t-test for independent samples was used. table 4 represents the results.

ISSN 2055-0219(Print), ISSN 2055-0227(online)

Table (4). Independent Samples T-test for gender performance

	t	df	Sig
Performance on MC	-2.22	31	0.03
Performance on Completion items	-1.11	31	0.276

As shown in table 4 it found that there are statistical significant differences in gender performance in MC test in favor of females, but the difference in gender performance was significant for completion items test (α=0.05).

The 3rd question: Do the psychometric characteristics vary by item formats?

Reliability of the 2 forms was checked using internal consistency (Cronbach α, Sperman-Brown coefficients and Guttman split half coefficient). table 5 represents these findings.

Table (5). Reliability

	Multiple choice	Completion
Cronbach α	0.50	0.88
Sperman- Brown coefficient	0.47	0.89
Guttman split half coefficient	0.46	0.87

As table 5 indicates, the completion item test was more reliable than multiple choice items test.

Item difficulties and item discriminations were calculated and table 6 represents the results.

Table (6). Items difficulties and discriminations

Item No.	difficulties		discriminations	
	MC	Completion	MC	Completion
1	.6364	.4545	0	.43
2	.5758	.0909	.85	.42
3	.8485	.3333	.57	.71
4	.5758	.6364	.3	1
5	.3636	.4242	.29	1
6	.4242	.0000	.42	0
7	.7576	.0606	.71	.15
8	.4242	.1212	0	.15
9	.0909	.3939	.32	.57
10	.5455	.4242	.43	1
11	.8485	.3030	.29	.85
12	.9091	.2727	.57	.85
13	.5152	.4242	.14	.29
14	.3333	.3939	.29	.57
15	.8182	.3030	.71	.57
16	.6667	.3030	.43	.85
17	.6061	.5455	0	.71
18	.9091	.2121	.71	.42
19	.3939	.4545	.43	.85
20	.8788	.3939	.43	.71
21	.7273	.3333	.57	1
22	.6061	.1818	.3	.42
Mean	0.61	0.32	0.40	0.61

As shown in table 6 the difficulty mean for MC items = 0.61 which indicates a medium difficult test, the difficulty for the items range was (0.1 – 0.9) and most of the items difficulty was

ISSN 2055-0219(Print), ISSN 2055-0227(online)

around (0.5) while the difficulty mean for completion items = 0.32 which indicates a difficult test, the item difficulty range was (0.0 – 0.64). The differences between item difficulties and item discriminations were significant (α=0.05) for the 2 forms as shown by independent sample t-test (table 7).

Table (7). Independent Samples T-test for item difficulties and discriminations

	t	df	sig
Item difficulties	5.07	42	0.00
Item discriminations	-264	42	0.01

DISCUSSION

The results of this study showed that students' performance was affected by items format, students performance on MC test was better than Completion; this could be resulted from the nature of the 2 forms, students have to recall information and to use higher order skills in order to answer completion items, this result is similar to the finding of Phipps and Brackbill (2009). which found that items that need more detailed information are more difficult. Female students achieved higher than males in 2 test forms, although the difference in means was not significant in completion item test and this could resulted from cultural perspectives, female students' in Jordan have to stay at home due to traditions after the university time, whereas; the male students can spend time with friends out of home which makes female students have much time for studying than, this result countered the findings of Hudson (2012) and Demas (1998).. This result meets with the findings of Culligan (2015) and Simbak, Aung, Ismail, Joush, Ali, Yaseen, Haque and Rebuan (2014). The results of this study also found that there were significantly differences in item formats characteristics (item difficulty, item discrimination and reliability).

Implication to Research and Practice

The accurate assessment of the students' achievement is considered to be the main issue in assessing students' performance; which depends upon the construction of the test. Test construction must match the students' ability and content. This study highlighted the relationship between items formats and their psychometric characteristics, and their influence upon students' achievement. So educators should take into their account the effect of item format in assessing the student performance.

CONCLUSION

The study recommended that faculties have to use different types of items format when they construct tests in order to match the students' preferences and abilities. Faculties have to inform their students about the test item formats; because each format needs special learning style. Decision makers at academic institutions need to motivate male students to fill the gap between males and females performance.

ISSN 2055-0219(Print), ISSN 2055-0227(online)

REFERENCES

Allam, S. (2007). *Measurement and Evaluation in teaching process*. Amman: Dar Almasira.

Alzude, N & Elian, H.(2005). *Principles of measurement & Evaluation in Education*. Amman: Dar Alfkr.

Center for innovations in teaching & learning improving your test questions, Retrieved from http:// www.cte.illinois.edu/testing/ exam/test_ques2.html.reterived in 21 Feb 2016.

Culligan, B.(2015). A comparison of three test formats to assess word difficulty. *Language Testing*. 32(4).503-520.

Demas, C(1998). Gender differences in , mathematics & science on a high school proficiency Exam: The role of Response format. *Applied Measurement in Education*. 11(3). 279-299.

Fulcher, G. & Davidson, F. (2007). *Language Testing & Assessment: an Advanced resource book*, New York: Routledge.

Harris, D. (1969). *Testing English as a Second Language*. New York: Mc Graw Hill in Haynie, W. (1994). Effect of Multiple – Choice & short Answer Test on Delayed Retention Learning. *Journal of Technology Education*. 6(1). 32-44.

Kufahi, T.(2003). *Measurement & Evaluation in Special Education*. Amman: Dar Almasira.

Mobalegh,A. & Barati, H. (2012). Multiple True–False & Multiple– Choice test Formats: A Comparison between Two versions of the same Test Paper of Iranian NuEE. *Journal of Language Teaching & Research*. 3(5). 1027-1037.

Phipps, S. & Brackbill, M. (2009). Relationship between Assessment item Format & Item performance Characteristics. *American journal of Pharmaceutical Education*. 73 (8).

Simbak, N. Aung, M. Ismail, S. joush, N. Ali, T. Yaseein, W. Haque, M. & Rebuan, H. (2014). Comparative study of different formats of MCQs: Multiple true-false & single best answer test formats, in a new medical school of Malaysia. *International Medical Journal*. 21(6). 562-566 .

Thorndike, R. & Hagen, E.(1986). *Measurement & Evaluation in Psychology & Education*. London: MacMillan.

ISSN 2055-0219(Print), ISSN 2055-0227(online)

APPENDIX

Test Form (A)

This test composed of 22 items, each item has (4) alternatives, circle the correct answer.

1- The variable that could be transformed into interval level of measurement is:
   a- Nominal
   b- Ordinal
   c- Interval
   d- Ratio

2- When a student had zero in a test this means:
   a- He/she did not know anything
   b- Not clever
   c- Did not answer any question
   d- His knowledge = zero

- A biology test composed of (50) MC items. the ratio allocated for Ecology unit = 30%, the ratio allocated for higher order skills = 40% and for knowledge and understanding = 60%, use the above information to answer the following 2 questions.

3- The number of items for Ecology unit which assesses higher order skills =
   a- 12
   b- 6
   c- 40
   d- 50

4- The ratio (%) allocated for Ecology knowledge and understanding =
   a- 12
   b- 6
   c- 40
   d- 50

5- If X= 2Y, then the level of measurement is :
   a- Nominal
   b- Ordinal
   c- Interval
   d- Ratio

6- The process by which intelligence is transformed to number (IQ) is called:
   a- Measurement
   b- Evaluation
   c- Assessment
   d- Diagnosis

7- Three students had the following marks in a test (25,19,18) the measurement level is:
   a- Nominal
   b- Ordinal
   c- Interval
   d- Ratio

ISSN 2055-0219(Print), ISSN 2055-0227(online)

8- The assessment characteristic in which parents, teachers, principal and students were involved in assessment process is called :
   a- Diagnostics
   b- Comprehensive
   c- Continuity
   d- Cooperation

9- The tests were classified in to power tests or speed test according to……
   a- Measurement field
   b- Test objectives
   c- Items format
   d- Test duration time

10- The examinee in one of these options express his/ her emotional feelings :
   a- Physics test items for 10^{th} grade students
   b- Check list used by teacher to evaluate students work in biology lab.
   c- Creative thinking test to explain scientific phenomena
   d- Check list used to indicate the prefered activities for the student during science lessons.

11- The comprehensive exam for diploma is considered as :
   a- Formative assessment
   b- Summative assessment
   c- Diagnostic assessment
   d- Measurement

12- "Students have to correct the mistakes in a paragraph". The objective level according to Blooms' taxonomy is:
   a- Understanding
   b- Application
   c- Analysis
   d- Synthesis

13- Class room discussion is considered as :
   a- Formative assessment
   b- Diagnostic assessment
   c- Summative assessment
   d- Measurement

14- Toleen is higher than 80% of the class students' the type of test is that indicate the above statement is:
   a- Norm-referenced
   b- Criterion referenced
   c- Oral referenced
   d- Can not be determined

15- " Median is used with ordinal scale measurement" this statement is :
   a- True
   b- False

ISSN 2055-0219(Print), ISSN 2055-0227(online)

16- When the students' performance in practical test is explained according to cut- off score, the type of test is :
   a- Norm-referenced
   b- Criterion referenced
   c- Standardized
   d- Practical

17- According to Language, the test you answering it now considered as:
   a- Standardized
   b- Norm-referenced
   c- Criterion referenced
   d- Practical

18- Measurement error could be found in :
   a- Instrument
   b- Person
   c- Trait
   d- All above

19- The type of assessment that used to determine the pre requests needed for teaching the new content is called :
   a- Formative
   b- Summative
   c- Standardized
   d- Diagnostic

20- The test that had a definite psychometric characteristics and had norms and piloted at a large sample is called:
   a- ICDL
   b- Standardized
   c- Tawjihi
   d- Driving license

21- "students has to write the dates of certain events". The objective level according to Blooms' taxonomy :
   a- Knowledge
   b- Understanding
   c- Application
   d- Analysis

22- When the student score affected by his gender, the type of test is called :
   a- Subjective
   b- Objective
   c- Norm-referenced
   d- Practical

Test Form (B)

This test composed of 22 items, each item has (4) alternatives, circle the correct answer.

ISSN 2055-0219(Print), ISSN 2055-0227(online)

1- The variable that could be transformed into interval level of measurement is……..

2- When a student had zero in a test this means…….

- A biology test composed of (50) MC items. the ratio allocated for Ecology unit = 30%, the ratio allocated for higher order skills = 40% and for knowledge and understanding = 60%, use the above information to answer the following 2 questions.

3- The number of items for Ecology unit which assesses higher order skills=……...

4- The ratio (%) allocated for Ecology knowledge and understanding = ……

5- If X= 2Y, then the level of measurement is …….

6- The process by which intelligence is transformed to number (IQ) is called…….

7- Three students had the following marks in a test (25,19,18) the measurement level is……..

8- The assessment characteristic in which parents, teachers, principal and students were involved in assessment process is called ……

9- The tests were classified in to power tests or speed test according to……

10- Write an example of a test in which the student express his/ her emotional feelings ……..

11- The comprehensive exam for diploma is considered as ……

12- " students have to correct the mistakes in a paragraph". The objective level according to Blooms' taxonomy is…….

13- The type of Assessment which depends upon class room discussion is considered as …….

14- Toleen is higher than 80% of the students' in her class, according to the test results interpretation the type of test is …….

15- " Median is used with ordinal scale measurement" this statement is true or false……...

16- When the students' performance in practical test is explained according to cut- off score, the type of test is ……..

17- According to Language, the test you answering it now considered as……

ISSN 2055-0219(Print), ISSN 2055-0227(online)

18- Measurement error could be resulted from …….

19- The type of assessment that used to determine the pre requests needed for teaching the new content is called ……

20- The test that had a definite psychometric characteristics and had norms and piloted at a large sample is called……..

21- "students has to write the dates of certain events". The objective level according to Blooms' taxonomy is ……

22- When the student score affected by his gender, the type of test is called……..

ISSN 2055-0219(Print), ISSN 2055-0227(online)