International Journal of Applied Linguistics & English Literature
ISSN 2200-3592 (Print), ISSN 2200-3452 (Online)
Vol. 3 No. 2; March 2014

Simple Versus Complex Stems in Multiple-choice Tests and their Effects on Students' Performance

Nader Zarza (Corresponding author)
Department of Language and literacy Education, Faculty of Education, University of Malaya
Tel: 0060176592543 E-mail: Naderzarza@yahoo.com

Nabeel Abedalazeez
Department of Educational Psychology and Counselling, Faculty of Education, University of Malaya
Tel: 0060162501958 E-mail: nabilabedalaziz@yahoo.com

Received: 24-11-2013 Accepted: 02-01-2014 Published: 01-03-2014
doi:10.7575/aiac.ijalel.v.3n.2p.237 URL: http://dx.doi.org/10.7575/aiac.ijalel.v.3n.2p.237

Abstract

Multiple-choice tests are used to assess learning outcomes in variety of disciplines. They are comprised from two main parts: stem and options (distractors). Stem, by conveying information and fore heading the problem to be answered plays a vital role in any multiple-choice test. Although there are large bodies of studies to design options in these kinds of tests, constructing stem items has received too little attention. Accordingly, much of current debate in constructing test items revolves around designing and building proper multiple-choice test options. This study tries to investigate the effect of item stem structure on students' performance in order to choose structure in designing stems in favor of students' high performance.

Keywords: Multiple-choice test, item-writing, stem

1. Introduction

Multiple-choice tests, as the most common used form of tests, play a significant role in evaluating progress and making decision for allocation learners to a higher level of learning in a small scale such as class level or in a larger scale such as graduation, promotion, certification, licensure, or placement (Haladyna, 2004). Accordingly, constructing multiple-choice question (MCQ) items requires more attention to obtain more reliable output for any decision. In large body of literature in regard of item writing, distractors or options have received a lot of attention comparing with item stems.

Putting a step forward in designing sound and reliable items and responding to a call for empirical studies of MCQ item construction, this study, by comparing structurally simple and complex stems, attempts to empirically investigate which structure of stem impacts students more effectively to achieve the best performance doing their MCQ exams. This comparative experimental study is expected to contribute and fill partly the gap in constructing the stem of MCQ tests. It hopes to enable teachers or test makers to design more reliable research based stem for their test. Moreover, this study intended to refine or confirm existing rules in writing well-defined item stems and finally, bring clarity to a set of confusing contradicted research findings in this regard. To meet these objectives, the study tries to answer this question:

- Is there a significant difference between students' performance in simple and complex stem of multiple-choice tests?

2. Literature Review

Harris (1969) was the first prominent researcher who brought the importance of test item constructing, item writing, to light by declaring that "general principal in testing is to confine the comprehension problems to either the lead stem or the options…, but not to insert problem in both. Accordingly, item writing became one important issue for researchers and studies in defining best items for test construction started. Generally, there are two main arguments resulted from various studies which are contradicted and each has their own reasons. First group of these studies maintain that the practice of inserting the comprehension problem in the stem should be avoided. Researchers such as Ascolon, Meyers, Davies and Smiths (2007), Collins (2006), Gronlund and Linn (2000), Haladyna and Downing (1989a), Khodadady (1999), Osterlind (2005), Passmore, Dobbie, Parchman, and Tysinger (2002) were against putting the comprehension problem in the stem of MCQ tests. They were in favor of designing stems with only necessary information and keeping it as short as possible. On the other hand, the second group's discussions are against the reviewed studies mentioned before. They have maintained that providing more information has significant effects on understanding items by students. Heaton (1988, p.56) on the face of the idea of using short and easy sentences, stated that "simple and short structure of stem cannot provide enough information to make stem understandable". Alongside, it was stated by many researchers that providing different structures of sentences in designing stem enhance students' comprehension (Chiang and Dunkel, 1992, cited in Ying-hui, 2007, p. 10; Gorjian, Jalilifar, and Mousavi , 2009; Parker and Chaudron,1987). It not only will increase test difficulty, but also will make item stem more comprehensible. Overall, the existing ideas in regard of choosing a definite structure for item stem, fail to resolve the contradiction in research findings to emphasize choosing a special structure based on a comparative study. Nearly all the studies dealt with the simple and complex stems separately while this study performed a comparative study between these two kinds of stems.

3. Methodology

3.1 Research design

The main object of this study was to determine whether simple or complex sentence stem affects students' performance in an administered multiple-choice test. In order to answer the question of interest in this research and because of the existence of a comparative problem, the design of this research was a true comparative experimental design solution. In this design, there were two equivalent groups, equal in English knowledge ability. In the first stage of the research, one group took the simple stem version of the multiple-choice test and the other group, the complex version. Moreover, in the second step, the simple group took the complex stem and the complex group took the simple stem multiple-choice test. To control the effect of students' performance in the proficiency test on students' performance on simple and complex, covariance analysis was used. Covariance is a measure of how much two variables change together and how strong the relationship is between them. It has a number of purposes but the two that are, perhaps, of most importance are: -to increase the precision of comparisons between groups by accounting to variation on important prognostic variables; - to "adjust" comparisons between groups for imbalances in important prognostic variables between these groups.

3.2 The Instrumentation

A sample proficiency 30-item test was administered to the population in defining homogeneous sample subject. Four versions of a 40-item achievement multiple-choice test from the IELTS preparation test books at the proficiency level. Two versions with all stems structurally simple and the other two versions were structurally complex stems. Version 1 of simple stem and version 1 of complex stems were for the same reading texts. And version 2 of simple stem and version 2 of complex reading text were for the same reading text. In both pair versions, just the stems were in the forms of complex and simple structure and the alternatives were the same. Two hundred graduate students from the University of Malaya were administered the proficiency test. Eighty-five students who were equivalent in their ability, recognized and separated by the proficiency test. They were considered as sample subjects. Afterwards all 85 individuals on the list were assigned consecutive numbers from 1 to 85. To have two equal groups to be randomly chosen and administered, arbitrary numbers were selected in the table till attaining two 40-member groups. One group of 40 members was named simple and the other 45 members, complex. Both proficiency test and main tests (simple and complex) received all the processes of refining. According to experts' revision and advice, the researcher revised the items and some were discarded. Next, the primary versions of the test booklets were administered to 60 students for item-trying out, item-correction, checking the time required for answering the tests, finding out test reliability, and checking item analysis. The range of item difficulty was from 0.21-0.85, whereas the discrimination of the item ranged from 0.18 - 0.65. Ten questions from each version were discarded due to their lowest discrimination indices. The Cronbach alpha for the test was .83. Accordingly, the final versions of the main tests were 40 items each and 30 items for proficiency test.

3.3 Sampling

Administering a 30-item sample English language proficiency test to the target population to choose sample subject. Both simple and complex stems for this multiple-choice test were included.
Selecting eighty-five homogeneous students who are equivalent in their ability, these students were randomly divided into two groups (one simple group and the other complex group).

3.4 administering the tests

In order to find out the impact of simple and complex structure- stem of multiple-choice tests to the sample subjects, tests were administered in two phases. The first version of 40-item multiple-choice tests with complex stems were administered to the complex group while 40-item of multiple-choice tests with simple stems administered to the simple group. In the second phase of administering the tests, the simple group was administered the test with complex stems while the complex group took the test with simple stems in order to avoid the students' getting used to one type of test and its effect on the result of the tests and also to observe a group achievement on different kinds of tests. The interval between each period was three weeks.

3.5 Data Analysis

The data obtained for this study was analyzed by using t-test, covariance, and Pearson correlation co-efficient through the Statistical Package for the Social Sciences (SPSS). This study used analyses of covariance in which student's level of proficiency used as a covariate variable. The dichotomies of students' answers were: "1" for correct answer and "0" for wrong answers. The total possible marks were 30 for the proficiency test and 40 for the main test.

4. Findings

To check the equivalence of the two groups, with respect to their performance on proficiency test, t-test was used. Table (1) shows the means and standard deviation and summary results of the t-test.

Table 1. Summary results of *t*-test for equity of Means

Stem	Number	Mean	Standard Deviation	*t*-value	*p*-value
complex	45	65.30	16.95	.191	.849
simple	40	65.78	14.25		

Table (1) shows that there is no significant difference between the performance of two groups (simple and complex) on proficiency test. Accordingly, the two groups are equivalent.

(t (85) = 0.849 , p > .05

To explore the differences between students' score in the simple and complex version, ANCOVA analysis was used. The assumptions of ANCOVA fulfilled. Table 2 shows the summary results of ANCOVA analysis.

Table 2. Summary results of covariance analysis

Source	Type III Sum of Squares	Df	Mean Square	F	Sig.	Noncent. Parameter	Observed Power(a)
Corrected Model	5508.698(b)	2	2754.349	20.495	.000	40.990	1.000
Intercept	3019.153	1	3019.153	22.465	.000	22.465	.997
PROFIC	4901.315	1	4901.315	36.470	.000	36.470	1.000
STEM	681.506	1	681.506	5.071	.027	5.071	.605
Error	11020.196	82	134.393				
Total	298193.000	85					
Corrected Total	16528.894	84					

Table (2) shows that there is a significant difference between the two groups (simple& complex) F(1, 82) = 5. 071 , P< 0.05. Student in simple stem (M =65.78, SD =14.25) scored higher than students in complex stem (M = 65.30 , SD =16.95).

To compare between the performance of the group one (complex version1, and the simple stem version 2) the t-paired independent sample was used. Table 3 shows the summary results of t-paired independent test.

Table 3. Summary of results of t-paired independent test

	Mean	***SD***	***t*-value**	***p*-value**
Complex version1	55.04	14.52	-2.601	.013
Simple version2	62.18	13.42		

Table (3) shows that there is a significant difference between the performance of group one on their performance on complex version 1 and simple version two. [t (45)= -2.601, p < .05]. The group one students scored higher in simple stems.

To compare between the performance of the group two (complex version2, and the simple stem version 1) the *t*-paired independent sample was used. Table(4) shows the summary results of t-paired independent test.

Table 4. Results of *t*-test comparison between simple version 1 and complex version 2

	Mean	***SD***	***t*-value**	***p*-value**
Simple version 1	60.40	13.06	2.26	.030
Complex version 2	53.67	14.17		

Table (4) shows there is a significant difference between simple stem test version 1 and complex stem test version 2. [t (40)=2.26, p < .05]. Group two students scored higher in simple stems.

To explore the relation between students' performance on complex stem and students' performance on the proficiency test, Pearson Correlation test was used. The resulting Pearson correlation co-efficient revealed that there is a significant relation between complex stem and proficiency test. [r (85) = 0.682, p < .05].

To explore the relation between students' performance on simple stem and students' performance on the proficiency test, Pearson Correlation Co-efficient was used. Results from Pearson Correlation Co-efficient revealed that there is a significant relation between simple stem and proficiency test. [r (85)= 58.7, p < .05].

5. Discussion

The stem is the foundation item of any MCQ test. After reading the stem, the student should know exactly what the problem is and what he or she is expected to do to solve it. If the student has to infer what the problem is, the item will likely measure the student's ability to draw inferences from vague descriptions rather than his or her achievement of the course objective. Hence those who are creating MCQ items must place such material in the stem to decrease the reading burden and more clearly define the problem in the stem. The findings of this study supported Haladyna and Downing's (1989a) claim that simple stems are necessary parts of MC item construction, and Khodadady's (1999) statement that "in the case of having extraneous clues in the stem, this very principle is violated and test makers should include just the context that is directly related to the keyed response".

Moreover, the findings of the research done by Ascalon, Meyers, Davies and Smits (2007) on the format of the item stem and its effect on item difficulty showed that the effect of stem is minimal. Their study is in agreement with the results of the present study that showed little variation in the effect of simple and complex stems on the students' performance. Additionally, the findings of this study agreed with the results of the study conducted by Passmore, et.al. (2002) claiming that acceptable stems are short (i.e., shorter than 20 words). It is also in agreement with the Osterlind's (2005) and Collins's (2006) findings. They concluded that the stem should have only the necessary information and it should be kept as short as possible.

The higher mean of the simple tests also showed that the simple group had better performance on the comprehension of the items. This might be because of cognitive extension that the simple stems give to the candidates. This seems to support the findings of Chiang and Dunkel (1992) who found simplicity does play a significant role in test comprehension. However, the findings of this study did not support the arguments made by a number of researches. The results of the study does not support Heaton's (1988, p. 56) findings asserting that simple sentence stems do not provide enough contexts and too little contexts are insufficient to establish any meaningful situation. The high mean of the complex tests rejected Gronlund and Linn's (2000) finding, claiming that the excessive length can confuse or distract candidates.

In spite of the fact that the mean score obtained from the tests with simple stems were just slightly higher than the mean score obtained from the tests with complex stems, the result of the basic statistics showed no significant difference between these two kinds of stems. This may be due to the unfamiliarity of the students with the complex stems, the similarity of the strategies that the students use in answering both kinds of tests, or the techniques used by the teachers.

6. Conclusion

Although the result of the present study revealed no significant difference between simple and complex stems, the mean scores obtained through basic statistics showed that the performance of the students taking the tests with simple stems was slightly higher than those taking tests with complex stems. This can be seen in the four versions of test administration in both groups during the treatment period. The results of this study did not fully conform to the studies that were in favor of the complex stems. It suggested that the stem should be written in the simplest, clearest and unambiguous way to avoid it being a reading test. In short, extra information in designing the stems of multiple-choice items does not guarantee the enhancement of test takers' performance significantly. Lengthy or very short stems do have their own disadvantages. While the former may make the learner confused or bored, the latter may provide the test takers (students) with incomprehensible data. The results of the present study showed that both simple and complex stems attracted somehow the same mean scores. However, it can be concluded that there is a need to focus on the amount of relevant and enough information in the design of the stem in a moderate status (i.e., not much or less information within the stem). This may facilitate students' performance in taking multiple-choice tests.

References

Ascalon, M. E., Meyers, L. S., Davis, B. W., & Smits, N. (2007). Distractor similarity and item-stem structure: Effects on item difficulty. *Applied Measurement in Education, 20*(2), 153-170.

Chiang, C. S., & Dunkel, P. (1992). The effect of speech modification, prior knowledge, and listening proficiency on EFL lecture learning. *TESOL quarterly, 26*(2), 345-374.

Collins, J. (2006). Education techniques for lifelong learning - Writing multiple-choice questions for continuing medical education activities and self-assessment modules. *Radiographics*, *26*(2), 543.

Gronlund, N. E., & Linn, R. L. (2000). *Measurement and evaluation in teaching*: Macmillan New York.

Haladyna, T. M. (1999). Developing and validating multiple-choice test items: L. Erlbaum Associates (Mahwah, NJ).

Haladyna, T. M., & Downing, S. M. (1989a). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2*(1), 37-50.

Harris, D. P., & Palmer, L. (1970). *A comprehensive English language test for speakers of English as a second language: listening*: McGraw-Hill.

Heaton, J. B. (1988). *Writing English language test*. New York, NY: Longman.

Khodadady, E. (1999). *Multiple choice stems in testing*: *Practice and theory*. Tehran, Iran: Rahnama.

Osterlind, S. J. (2005). Creating quality multiple choice questions. Retrieved from http://www.ecdledg.com/news/Edvoece-arc-05.10-06.learn-34k

Parker, K., & Chaudron, C. (1987). The effects of linguistic simplifications and elaborative modifications on L2 comprehension. *UHWPESL,* 6(2), 1.

Passmore, C., Dobbie, A. E., Parchman, M., & Tysinger, J. (2002). Guidelines for constructing a survey. Research Series, 34(4), 281-286. Retrieved June 17, 2006, from www.stfm.org/fmhub/fm2002/apro2/rs1.pdf.

Ying-hui, H. (2007). An investigation into the task features affecting EFL listening comprehension test performance. *Asian EFL Journal, 8*(2), 1-15. Retrieved December 19, 2007, from http://www.asian-edf-journal.com/june_06_hyh.php_54k.