A Comparison between Three-and Four-Option Multiple Choice Questions

Afsaneh Dehnad\textsuperscript{a}, Hayedeh Nasser\textsuperscript{b,*}, Agha Fatemeh Hosseini\textsuperscript{c}

\textsuperscript{a}School of Health Management and Information Sciences, Iran University of Medical Sciences, Tehran, Iran.
\textsuperscript{b}School of Medicine, Iran University of Medical Sciences, Tehran, Iran
\textsuperscript{c}Department of Math and Statistics, School of Health Management and Information Sciences, Iran University of Medical Sciences, Tehran, Iran.

Abstract

To write plausible options for multiple choice questions (MCQs) is often a demanding job for test makers. Three-option MCQs might be a solution for the process of test making as well as test taking. This study was an attempt to compare three-option MCQs with four-option in terms of test usefulness. The participants were 114 second semester medical students taking the tests as their final examination concurrently while the time of test completion was recorded. The collected data were analysed by SPSS with the use of an independent t-test. The findings showed some differences in usefulness of these two test formats.

Keywords: Multiple Choice Questions; test usefulness; three-option multiple choice; four-option multiple choice

1. Introduction

Multiple choice questions (MCQs) enjoy widespread popularity owing to particular features which make them highly objective instruments for measurement. Despite their long history which dates back to the early 1900s (Landrum, 1993), these tests are still in extensive use in higher education for their high reliability, high content coverage, rapid and economical scoring and openness to item analysis. On the other hand, they have a negative reputation as having low validity, testing factual knowledge rather than a high level of cognitive knowledge,

\* Corresponding author. Tel.: +982126121178; fax: +982126121186.
E-mail address: hnasser60@yahoo.com
increasing guessing (especially as the number of options decreases) and consuming more time in the process of test development in particular finding plausible distractors. As the demands of multiple choice tests entail a high level of item writing skills and experience (Hughes, 2003), writing meaningful, precise items which leads to fair scores is one of the main concerns of language teachers. It is an art, Ebel (1951) notes, that can be acquired through practice and experience; however, two groups of teachers are more at the risk of writing poor items: novice teachers and teachers who may lack adequate competence in test writing. These teachers are often engaged in making multiple choice tests for diagnostic purposes in classroom-based assessment. Many times the four-option items they write appear to be defective containing implausible options. Look at the following example taken from a test made by a novice teacher:

We must protect and preserve our ……resources for the next generation that has to encounter scarcity of natural resources.

a. moral  b. vital  c. oral  d. arsenal

The example shows that the fourth option is not plausible; therefore, the likelihood of omitting the item or searching for another item increases, the process of item writing, in turn, becomes longer and more difficult, and the test maker may use options like “all of the above” or “none of the above” as the last resorts, again increasing the chance of guessing.

The literature shows many studies in favor of three-option multiple choice tests. Landrum et al (1993) found that students’ performance on a three-option was better than a four-option test format. In a study of the comparison of three, four and five option MCQs, Farhady & Shakery (2000) found no significant difference in psychometric characteristics of these test formats, thereby recommending three-option MCQs as a better test format than four or five options. Rodriguez (2005) in a meta-analysis of 80 years multiple choice questions found that three-option tests improve content coverage while not affecting the psychometric quality of the tests. Vyas and Supe (2008), through a literature review, found no significant difference between three, four and five option test formats for assessing medical students. However, three option tests improved efficiency and administration as fewer distracters were needed, thus saving time and space for inclusion of more items and more content, while requiring less reading time for test takers. Tarrant and Ware (2010) comparing psychometric properties of three- and four-option test formats in assessing nursing students, supported the three-option test format for containing more functioning distracters and higher discriminating effects, while requiring less time for test development and administration and providing no advantages in reliability and validity. In a study of three, four, and five options for a listening test Lee and Winke (2012) found that the three-option multiple choice test appeared to have a higher mean difference than the other test formats, while there was no difference in terms of item discrimination. Delgado and Prieto (2012) provided evidence favoring three-option test formats as they found no decrease regarding item discrimination and tests reliability as compared with four-option test format in computerized examinations.

Although the literature is in favor of three-option multiple choice, four-option multiple choice is conventionally used in many medical schools where the dominant trend is toward objective tests. This paper was an attempt to examine three and four-option multiple choice test to provide more evidence for three-option multiple choice test from the perspective of test usefulness, proposed by Bachman and Palmer (1996), as the most important quality of a test. Test usefulness, consisting of several functional variables: reliability, validity, authenticity, instructiveness, impact, and practicality, is a tool which enables the test developer to evaluate the test he designs.

2. Method

This comparative-descriptive study was conducted in the school of medicine at Iran University of Medical Sciences(IUMS) in 2013.
2.1. Participants

The participants of the study were 114 second semester students of medicine who were taking technical English language 1. There were 74 females and 40 males with the age range of 19-23 (Table 1). On the basis of their mid-term exam grades (Mean= 3), the students were divided into two groups of pre-intermediate (PI), those who scored below the mean, and intermediate (I), those whose scores were above the mean. Each of these groups was then divided into two subgroups: the groups taking the three-option MCQs (I3 and PI3) and the groups taking four-option multiple choice test (I4 and PI4). Table 1 shows the number of students in each subgroup.

<table>
<thead>
<tr>
<th>groups</th>
<th>Total number</th>
<th>females</th>
<th>males</th>
</tr>
</thead>
<tbody>
<tr>
<td>I3</td>
<td>34</td>
<td>15(44.1%)</td>
<td>19(55.9%)</td>
</tr>
<tr>
<td>I4</td>
<td>33</td>
<td>19 (57.6%)</td>
<td>14(42.4%)</td>
</tr>
<tr>
<td>PI3</td>
<td>23</td>
<td>8 (34.8%)</td>
<td>15(65.2%)</td>
</tr>
<tr>
<td>PI4</td>
<td>24</td>
<td>12 (50%)</td>
<td>12(50%)</td>
</tr>
</tbody>
</table>

2.2. Test characteristics

The test was an achievement test with the conventional format of the final exam in the medical school at IUMS. It consisted of four parts of vocabulary, grammar, reading comprehension and medical terminology (Table 2).

<table>
<thead>
<tr>
<th>Test section</th>
<th>Number of items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vocabulary</td>
<td>20</td>
</tr>
<tr>
<td>Grammar</td>
<td>6</td>
</tr>
<tr>
<td>Reading</td>
<td>14</td>
</tr>
<tr>
<td>Medical Terminology</td>
<td>40</td>
</tr>
</tbody>
</table>

All the items were initially written in four-option MCQs for the first version of the test to be administered as four-option MSQs. Then, the four-option test format was changed into three-option according to the subjective judgment of the course instructors who eliminated the least plausible option per item to be administered as three-option MCQs as the second version of the test. Each item was worth one point for both versions and there was no deduction for negative marking.

2.3. Statistical procedure

The items were checked for face and content validity, coverage, and authenticity by three ESP instructors. The reliability of both test versions was measured with Chronbach’s alpha inter-reliability item ($r_3 = 0.86$ and $r_4 = 0.85$, respectively). Both versions of the tests were administered as the final examination for all four groups (PI3, PI4, I3 and I4) concurrently while the researchers were present to record the time of test completion. Kolmogrov-Smirnov test was performed to check the normality of the distribution. Descriptive
statistics was used to check the effect of sex and age, and an independent t-test was applied to compare the means of the two groups.

3. Results and discussion

The mean score of the three-option MCQs was higher than four-option MCQs in the intermediate group; however, this difference was not statistically significant (P= 0.061). This finding suggests that students with higher language ability could easily select the correct answer by their language knowledge, strategies and visual processing whether it be a four-option or three-option MCQs. The mean score of three-option MCQs was higher than four-option in the pre-intermediate group and the difference was significant (p= 0.045), i.e. the pre-intermediate group performed better in three option version of the test. This finding indicates that less proficient students are more dependent on word meaning, and reading is a more bottom-up process of information whereas it is a top-down process of reading and grasping the information for more proficient students (Carrell, 1991). As for the findings of this study, the less proficient students had to spend more time on processing the information and finding the correct answer which were time consuming in the case of four-option MCQs but led to a better performance in three-option MCQs.

<table>
<thead>
<tr>
<th>Groups</th>
<th>Mean and standard deviation</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intermediate</td>
<td>I3: 70.12 ± 4.89</td>
<td>P= 0.06</td>
</tr>
<tr>
<td></td>
<td>I4: 67.73 ± 5.58</td>
<td></td>
</tr>
<tr>
<td>Pre-intermediate</td>
<td>PI3: 60.09 ± 7.15</td>
<td>P= 0.045</td>
</tr>
<tr>
<td></td>
<td>PI4: 55.83 ± 7.01</td>
<td></td>
</tr>
</tbody>
</table>

The results showed that the reliability of 3 option tests increased slightly. This finding is in line with the findings of Berrios et al. (2005) and Rodriguez (2005). Rodriguez, in his systematic review including studies from 1920-1990s, reported that in most cases, when the number of options is reduced, reliability decreases except in the case of a reduction from four to three-options where the reliability increases slightly. However, a few studies report no difference in reliability between three and four-option tests (Lee and Winke, 2012; Tarrant & Ware, 2010; Vyas & Supe, 2008; Delgado & Gerado, 1998). The inconsistency in the findings might be explained by variables like the method of option deletion (random vs. ineffective) emphasized by Rodriguez (2005), the number of items and other test specifications differing across studies.

The reduction in the number of options from four to three did not lead to significant differences in mean item difficulty (four-option MCQs= 0.79, three-option MCQs= 0.82). As for the item discrimination, the results showed that both versions could similarly discriminate intermediate from pre-intermediate students (four-option MCQs=26.02, three-option MCQs= 27.42, p< 0.001).

The validity of both versions was the same because there were two identical tests with the same number of items. Unlike Landrum, 1993 who reported an improvement in validity of the test items, the literature does not provide strong evidence supporting a difference in validity (Rodriguez, 2005; Vyas, Supe, 2008). Similarly, authenticity and inter-activeness of the three-option tests did not change as we used the same text, stem, and options for adjustments.

The test impact, as a variable of test usefulness, affects educational systems and individuals within those systems (Bachman and Palmer, 1996). In this study, the impact for PI3 group was favorable as they performed
better than their counterparts, PI4, suggesting that the scores were fairer for PI3 who might receive a positive feedback on their performance which in turn affects the decision made about them by the teacher. This was also discussed by Berrios et al. (2005) who came to the same findings. As for the test impact on educational systems, the more evidence that is provided for three-option MCQs, the more likely it is that they would be adopted in testing systems and the less bias there would be towards four-option MSQs.

The practicality of the test, which is how a test is designed and administered within the available resources, (Bachman & Palmer 1996) was tried and tested in this study. As for practicality, finding a plausible fourth option took time which otherwise could be spent for covering more content and more items. The present study supports the practicality of three-option MCQs in terms of ease of test development, more content coverage, more item inclusion and fewer faulty distracters. The practicality of three-option MCQs was also approved in former studies conducted by Farhady and Shakery, 2000; Berrios et al., 2005; Vyas and Supe, 2008; and Tarrant and Ware, 2009.

The findings on the time recorded for test completion showed no significant difference among the four groups of PI3, PI4, I3, and I4. This finding is in contrast with the findings of former studies (Landrum, 1993, Owen & Froman, 1987; Stratton & Catts, 1980 cited in Landrum, 1993; Berrios et al., 2005) arguing that three-option MCQs were less distracting to less proficient students as they could complete the test more quickly than the four-option. The finding of the present study, however, might be related to the less formal nature of achievement tests as opposed to large scale tests administered on a predetermined time and speed basis. In other words, achievement tests are not speeded and are often administered as a power test in which every student has enough time to complete the test. The findings of the present study support Grier’s (1976, cited in Rodriguez, 2005) assumption that time of test completion is affected by such factors as item complexity, as well as the number of both items and options.

Although PI3, PI4, I3, and I4 were homogenous and took the tests concurrently, which is an advantage over the former studies, the present study has some limitations. The design of the three-option test was based on subjective judgment of the test makers and this could have affected the findings of the study. Another limitation of the study was that the original test was designed by experienced ESP teachers and not by a novice teacher who due to inadequate competence of item writing might have developed a different test with different results.

4. Conclusions

The findings suggest that three-option MCQs save time for covering more content and items in the test, thereby increasing test validity and reliability. Three-option MCQs seem to be easier and less demanding for novice and inexperienced teachers as they would not be forced to use implausible and defective distracters in writing items. However, guessing is said to be a major problem as the number of options decreases. According to Bachman and Palmer (1996), there are two forms of guessing: informed guessing which requires partial knowledge of the subject matter and random guessing which is not based on content knowledge. As language teachers, informed guessing should be taken into account and students should be trained and encouraged to make informed guessing on the basis of their knowledge. Other solutions for decreasing the rate of guessing is to provide sufficient time for all the test takers to complete the test and match the difficulty level of the test with the ability level of the test takers. Three-option MCQs is recommended for novice teachers, classroom-based tests and achievement tests requiring more content to be covered in a short period of time whereas four-option MCQs could be more appropriate for high-stake tests or whenever it is functional or feasible.
Acknowledgements

We would like to thank our colleagues who helped us construct the initial form of the test and reviewed the tests for content and face validity. We are also thankful to our students as without their participation we would not be able to conduct this study.

References