Five-Option Vs. Four-Option Multiple-Choice Questions

B. Budiyono, *budiyono@ukwms.com,* UKWMS, Surabaya, Indonesia

Abstract: Multiple-choice questions (MCQs) may provide test takers with three, four, or five options and are appreciated for reliability and economic scoring. Five-option MCQs demand much more energy, experience, time, and expertise and may probably be considered to be more difficult four-option and three-option MCQs. Previous studies involved a great number of questions and participants. This study investigated the difference between five-option and four-option MCQs through the deletion of non-functioning distracters (NFDs) in proportion to a classroom-based test by administering 28 MCQs to two intact classes of 34 participants. The results show that there was a significant difference in participants' scores (p 0.030< 0.05), a significant difference in the number of NFDs (p 0.01<0.05), no significant difference in item facility (p 0.485>0.05), and significant difference in item discrimination (p 0.01<0.05). Classroom teachers are free to choose either the 5-option or 4-option version, depending on the purpose of the test.

Key words**:** five-option, four-option, non-functioning distractor

1. INTRODUCTION

A multiple-choice question (MCQ) consists of a stem as the question and a particular number (three, four, or five) alternatives or plausible options, one of which is the best or the right answer to the question in the stem. The ease in marking makes them preferred for large classes. A test creator adopts three-option MCQs because of the difficulty in constructing plausible distracters. Practically three-option MCQs are easier to construct than four-option and five-option MCQs. Any implausible distracter is referred to as test error and may unnecessarily require some extra time to complete reading the options. Three-option questions will require less time and result in more valid score results. This adoption is the improvement in test efficiency and administration with three-option questions (fewer distracters, less space, less reading time, and less construction time.

Multiple-choice questions are appreciated for a high level of reliability and widely used in higher education due to "high content coverage, rapid and economical scoring, and openness to item analysis (Dehnad, Nasser, & Hosseini, 2014). Constructing plausible options certainly requires experience and skills. It follows that the higher the number of options (e.g., five options), the more demanding the test construction will be. This belief will raise a question of the quality of the plausibility of the options whether these five options function in the test. Multiple-choice questions (MCQs) are sometimes in the format of one stem with five options, e.g., *UNAS* (Indonesian National Examination) and the GRE (Graduate Record Examination) but also in the format of one stem with four or three options. The question is whether the MCQs with a smaller number of options perform better than those with a greater number of options.

A problem that arises with MCQs is the requirement of plausible options. For example, providing five plausible options for each stem is certainly more difficult and time-consuming. These efforts do not guarantee that the 5-option MCQs will perform better than the 4-option counterpart MCQs. It is reasonable to provide fewer options. It takes less time to develop three options, and this makes it possible to increase the number of MCQs to defend against any potential decrease in reliability (Aamodt, and Shane, 1992) in (Scheneid et al.,2014).

Moreover, 17% more 3-option MCQs could be added by reducing 5 to 3 options per item (Owen & Froman, 1987) (Scheneid et al., 2014).

Previous studies involved a great number of MCQs and participants. Classroom-based assessment, however, would be conducted to intact classes that are usually small. Moreover, MCQs would be a portion of a classroom test that would involve short-answer questions and extended essay questions. For this reason, this study was conducted in intact classes with a small number of MCQs, as usually happens in a reading comprehension test.

Some previous studies are worth reviewing. Rahma et al., (2017) investigated the effects of reducing the number of options on an MCQs examination by administering forty MCQs, with one correct answer for each question, in two sets. In the first one, four options were given, including one key answer and three distractors. In the second set, one of the three distractors was deleted randomly, and the sequence of the questions was kept in the same order. Any distracter chosen by less than 5% of the students was regarded as non-functioning.

Kuder-Richardson Formula 20 (KR-20) was used to measure the internal consistency and reliability of an examination with an acceptable range of 0.8–1.0. A significant difference was observed in discrimination and difficulty indices for both sets of MCQs. More distractors were non-functional for set one (of four options) but slightly more reliable. The reliability (KR-20) was slightly higher for set one (of four options). The average marks in option three and four were 34.163 and 33.140, respectively. Their conclusion was that, in comparison to set 1 (four options), set 2 (of three options) was more discriminating and associated with low difficulty index, but its reliability was low.

Tarrant, Ware, & Mohammed (2009) investigated the proportion of non-functioning distractors on a sample of seven test papers administered to undergraduate nursing students. An MC test of 514 items with four options, including one correct answer to each question was given to 121 EFL university students. For this purpose, nonfunctioning options were defined as ones that were chosen by fewer than 5% of examinees and those with a positive option discrimination statistic as found in the item analysis. These were the results. The proportion of items containing 0, 1, 2, and 3 functioning distractors was 12.3%, 34.8%, 39.1%, and 13.8% respectively.

Overall, the items contained an average of 1.54 functioning distracters. Only 52.2% (n = 805) of all distractors were functioning effectively and 10.2% (n = 158) had a choice frequency of 0. Items with more functioning distractors were more difficult and more discriminating. The conclusion is that the low frequency of items with three functioning distractors in the four-option items suggests that teachers have difficulty developing plausible distractors for most MCQs. For them, test items should consist of as many options as is feasible as regards the item content and the number of plausible distracters.

Kilgour & Tayyaba (2016) explored non-functioning distractors in a sample of 480 five-option SBA (single best answer) questions of 327 university students. The study followed a two-step procedure, where the first step was to determine the frequency of non-functioning distractors across the sample of exam paper by analyzing the frequency of selection at the below 5 % level. It followed two steps: Firstly, for a reduction from five options per question to four, and secondly, for a reduction from five options to three. Overall, only 34 questions (7.1 %) of the 480 included in this study contained four functional distractors, while 92 (19.2 %) contained three. The greatest proportion of questions, 159 (33.1 %), had two functional distractors, while many questions contained only one (127, 26.5 %). Finally, 68 questions (14.2 %) contained no functional distractors and were therefore completely non-discriminatory. The analysis of the performance of the 1920 distractors reveals that 1062 (55.3 %) of the distractors were non-functional, with 341 of the distractors (17.8 %) being so implausible that they were never chosen. Of the 858 (44.6 %) functional distracters that were analyzed, only 206 (10.7 %) were chosen by more than 20 % of the examinee cohort. Testing

the changes of the original five-option model as a baseline, a series of paired-samples t-tests were carried out. For all years, the changes in difficulty between the five-option version of each paper and the four-option version, and the four-option version and the three-option version were all statistically significant. For the year three sample papers, discrimination was equivalent across the five-option and four-option exam models, but it decreased by a statistically significant degree for the three-option model. Discrimination then significantly decreased when the three-option model was employed, although discrimination for the three-option model remained significantly higher than for the original five-option version. Only 7.1 % of the questions contained four functional distractors, and 73.8 % of questions contained two or fewer.

Dehnad et al. (2014) compared the use of three and four options in the MCQs and found that the mean score of the three-option questions was slightly higher than that of the four-option MCQs. It was not statistically significant (P= 0.061) in the intermediate group. An interpretation of this finding is 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." On the contrary, the mean score of the three-option MCQs was significantly higher than that of the four-option MCQs (p= 0.045) in the pre-intermediate group. These lower-level students might have spent more time on processing the greater number of options.

Panczyk, M et al. (2014) investigated the effect of changing the number of options by administering 250 multiple-choice exam questions that consisted of 150 4-option items in 2009, 2010, and 2012, and 100 5-option items. The administration of 4-option MCQs in the years of 2009, 2010, 2012, resulted in the mean scores of 30.2, 25.6, and 31.5. The administration of 5-option MCQs in the years of 2011 and 2013 yielded the mean scores of 24.3, and 25.7. In addition to similarity in the mean scores, there was no significant difference in discrimination power (ANOVA test, P > 0.05). The conclusion is that the addition of options did not significantly improve the quality of the tests. For this reason, a test creator use of 4-option questions

The purpose of this study was to examine the effect of the reduction of the number of options, from five to four. The deletion of non-functioning distracters conducted it. More specifically, the purpose was to determine whether there as a significant difference in the participants' scores after taking the two versions, the 5-option and the 4-option MCQs.

Concerning the purpose of this study, two hypotheses were formulated. The null hypothesis is that there were no significant differences between the 5-option and 4-option versions in terms of the participants' scores, the number of NFDs, facility, and discrimination. The alternative hypothesis is that there was a significant difference.

2. METHOD

Thirty-four participants were attending the upper-intermediate reading class (Reading B) of the odd semester of AY 2017/2018. They took the 5-option multiple-choice reading test two weeks before the mid-semester test and the 4-option multiple-choice reading test two weeks before the end-of-semester test, i.e., nine weeks after the first test.

The instrument consisted of 2 paired tests taken from GRE. The first contained 28 multiple-choice reading comprehension test items with five options per item, and the second consisted of the same reading passages and the same number of items with four options per item. An item analysis was conducted to determine the non-functioning distracters. They referred to the distracters that were not chosen by any test taker or that were chosen by the smallest number of test-takers (≤5% of 34 test-takers). These distracters were deleted. The

deletion happened to the non-functioning distracters that appeared the first along the A-B-C-D-E sequence. This second test was the same as the first test minus one distracter.

The data were the participants' scores of the two versions of the test. The scores of the first test (5-option MCQs) were computed in item analysis in excel for establishing the facility degree, the discrimination power, the reliability degree (KR20), and for deleting the non-functioning distracters. The scores of the second version (4-option MCQs) were also computed in item analysis. The paired scores of the two versions, the paired scores of the facility degree, and the paired scores of the discrimination power were finally computed in SPSS.

3. RESULT

The number of non-functioning distracters was determined by calculating 5% or less of the number of the subjects, i.e., 5% of 34 is 1.7 rounded up to 2 or less, as presented below.

Table 4.1 Number of non-functioning distracters of 30 MCQs

Versions	NFDs	KR20	Facility	Discrimination
5-option	78 (55.71%)	0.74	0.79	0.58
4-option	59 (52.68%)	0.79	0.80	0.37

The table shows that the number of non-functioning distracters is bigger than a half of the number of options both in the 5-option MCQs (78 out of 140) and the 4-option MCQs (59 out of 112). This finding shows that the modification utilizing the deletion of NFDs fails to minimize the number of NFDs in the modified version. There may be a question of whether the 59 NFDs in the 4-option version include the NFDs in the 5-option version. There is at least one NFD in each 5-option MCQs, and the number of NFDs ranges from 1 to 4. This result also happens to the 4-option version, where the number of NFDs ranges from 1 to 3. This result should be a little away from expectation because, ideally, the number of NFDs in the 4-option version should have been 50 (78 minuses 28). This result is not in line with the finding by Hingorjo & Jaleel (2012) that 42% of 50 MCQs with five options were free from NFDs and the finding by Tarrant et al. (2009) that 12.3% of 514 MCQs questions with four options were free from NFDs. These findings are neither in line with the finding by Kilgour & Tayyaba (2016), who analyzed 480 questions with 1920 distracters and revealed that 34 questions (7.1%) performed with 4 FDs.

The table also informs the reliability level, the facility level, and the discrimination level. The interpretation of KR20 is that the two versions are reliably good for classroom tests. The discrimination indices describe that the 5-option version better is than the 4-option version and that the two versions are very good and good, respectively (Gronlund, N, 1981). The facility indices, however, suggest that the two versions are easy in general. The facility indices of the 5-option questions range from 0.62 to 0.97, and the facility indices of the 4-option questions range from 0.56 to 0.97, indicating that the items are of varying degrees of the facility.

An additional interpretation may be that there are no varying indices of the facility of individual questions, i.e., the questions of both versions belong to the easy level in MCQs. For this purpose, KR21 renders 0.72 for the 5-option version and 0.79 for the 4-option version to indicate that the scores are also reliably good for classroom tests.

The paired t-test was taken to determine whether there was a significant difference between the 34 test-takers' performance in both the 5-option and 4-option versions by computing their scores, as shown in table 4.2.1.

The results show that there is a significant difference between the test takers' scores in the two versions or that the scores of the second version are significantly higher than those of the first version, although there is no significant difference in terms of the facility.

Table 4.2.1 Paired t-Test of Test Takers' Scores

Paired Samples Test

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
Pair 1 data1 - data2	-.500	1.285	.220	-.948	-.052	-2.269	33	.030

The results show that there is a significant difference between the test takers' scores in the two versions or that the scores of the second version are significantly higher than those of the first version, although there is no significant difference in terms of the facility.

The following table provides the results of the t-test of the difference between the number of NFDs.

Table 4.2.2 Paired t-Test of NFDs

Paired Samples Test

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
Pair 1 NFD1 - NFD2	.679	.983	.186	.297	1.060	3.652	27	.001

The results indicate that there is a significant difference in terms of the number of NFDs. This finding should imply that the 4-option version might probably be preferred, although no question in this study is free from NFDs. There are other reasons for the preference of 4-option MCQs, e.g., "It is very difficult, even for a well-trained instructor, to provide many functioning distracters; otherwise he would just add distracters for completion" (Rahma et al., 2017) and "Writing plausible distracters is time consuming and the most difficult part of preparing MCQs" (Vyas & Supe, 2008). Shizuka et al., (2006) practically argued for three options for the sake of efficiency in "stationery, printing and test administration costs."

The table yields the results of the t-test of the difference in the facility level of the first version and second version.

Table 4.2.3 Paired t-Test of Facility

Paired Samples Test

	Paired Differences	t	df	Sig. (2-

		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				tailed)
					Lower	Upper			
Pair 1	Facility1 - Facility2	-.01429	.10786	.02038	-.05611	.02754	-.701	27	.489

It confirms that there is no significant difference or that the two versions are more or less of the same level of the facility.

Table 4.2.4 is about the difference between the first and second versions in terms of discrimination.

Table 4.2.4 Paired t-Test of Discrimination

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Discrimination1 – Discrimination	.21321	.42759	.08081	.04741	.37901	2.639	27	.014

The results mean that the first version is significantly more powerful than the second version in discriminating against the upper and lower achievers, although there is no significant difference in terms of the facility (The computation in excel also shows a significant upper-lower difference at p $0.01 \leq 0.05$). This result is also true with these previous studies. Reducing the number of options from five to four led to a small but significant reduction in reliability and a small decrease in discrimination power (Ramos & Stern, 1973) in (Thanyapa & Currie, 2014). A similar study is about the increase and the decrease of item discrimination power that resulted from the reduction of options from five to three options, and the greatest decrease of item discrimination happened when the options were reduced from five to two (Rodriguez, 2005).

4. CONCLUSION

Deletion of NFDs for the modification of 5-option to 4-option MCQs with a small number of MCQs and intact classes results in no significant difference in the facility. However, it yields significant differences in scores, number of NFDs, and discrimination.

These findings may suggest that classroom reading comprehension instructors may have their own decision regarding the version of MCQs to administer to their students. Quality 5-option MCQs such as GRE does not have to be reduced to a 4-option version when the purpose is especially to discriminate the upper-achievers from low-achievers. When there is a purpose to fill classes with a sense of achievement for a higher score, a test creator could do the deletion. This should not mean that 4-option MCQs may be made more discriminating by adding one option because this option may turn to be an NFD despite much energy and time.

It is noteworthy that item analysis data are tentative due to the influence of different factors, such as the number of participants and items. A retest would be useful to explore the relative usefulness of the items, e.g., when participants almost constantly neglect particular items, these items would have to be revised because of very low plausibility, or, when most of

the participants fail to select the right options, these items are probably too difficult.

Further research could investigate whether a 3-option multiple-choice test would be better in the sense that all the options are selected by more than 5% of the test takers. This test can be conducted by administering a 4-option test, analyzing the items, deleting one option that is selected by 5% or less of the test takers. These findings would probably confirm whether a 3-option multiple-choice test for classroom purposes is preferred and thereby might encourage classroom teachers to develop their tests with three options.

REFERENCES

Dehnad, A., Nasser, H., & Hosseini, A. F. (2014). A Comparison between Three-and Four-Option Multiple Choice Questions. *Procedia - Social and Behavioral Sciences*, *98*, 398–403. https://doi.org/10.1016/j.sbspro.2014.03.432.

Gronlund, N. (1981). *Measurement and Evaluation of English self in Teaching*. New York: Macmillan Publishing.

Hingorjo, M. R., & Jaleel, F. (2012). Analysis of one-best MCQs: The difficulty index, discrimination index, and distractor efficiency. *JPMA. The Journal of the Pakistan Medical Association*, *62*(2), 142–147.

Kilgour, J. M., & Tayyaba, S. (2016). An investigation into the optimal number of distractors in single-best answer exams. *Advances in Health Sciences Education*, *21*(3), 571–585. https://doi.org/10.1007/s10459-015-9652-7.

Panczyk, M et al. (2014). Comparison of four- and five-option multiple-choice questions in nursing entrance tests. *ICERI: Proceedings*.

Rahma, N. A. A., Shamad, M. M. A., Idris, M. E. A., Elfaki, O. A., Elfakey, W. E. M., & Salih, K. M. A. (2017). Comparison in the quality of distractors in three and four options type of multiple-choice questions. *Advances in Medical Education and Practice*, *8*, 287–291. https://doi.org/10.2147/AMEP.S128318.

Rodriguez, M. C. (2005). Three Options Are Optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years of Research. *Educational Measurement: Issues and Practice*, *24*(2), 3–13. https://doi.org/10.1111/j.1745-3992.2005.00006.x.

Schneid, S. D., Armour, C., Park, Y. S., Yudkowsky, R., & Bordage, G. (2014). Reducing the number of options on multiple-choice questions: Response time, psychometrics, and standard-setting. *Medical Education*, *48*(10), 1020–1027. https://doi.org/10.1111/medu.12525.

Shizuka, T., Takeuchi, O., Yashima, T., & Yoshizawa, K. (2006). A comparison of three- and four-option English tests for university entrance selection purposes in Japan. *Language Testing*, *23*(1), 35–57. https://doi.org/10.1191/0265532206lt319oa.

Tarrant, M., Ware, J., & Mohammed, A. M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: A descriptive analysis. *BMC Medical Education*, *9*(1), 40. https://doi.org/10.1186/1472-6920-9-40

Vyas, R., & Supe, A. (2008). Multiple choice questions: A literature review of the optimal number of options. *The National Medical Journal of India*, *21*(3), 130–133.