Title	A comparison of the psychometric properties of three- and four-option multiple-choice questions in nursing assessments
Author(s)	Tarrant, M; Ware, J
Citation	Nurse Education Today, 2010, v. 30 n. 6, p. 539-543
Issued Date	2010
URL	http://hdl.handle.net/10722/134855
Rights	NOTICE: this is the author's version of a work that was accepted for publication in Nurse Education Today. Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version was subsequently published in Nurse Education Today, 2010, v. 30 n. 6, p. 539-543. DOI: 10.1016/j.nedt.2009.11.002

Elsevier Editorial System(tm) for Nurse Education Today
Manuscript Draft

Manuscript Number: NET-D-09-00044R2

Title: A COMPARISON OF THE PSYCHOMETRIC PROPERTIES OF THREE- AND FOUR-OPTION MULTIPLE-CHOICE QUESTIONS IN NURSING ASSESSMENTS

Article Type: Full Length Article

Corresponding Author: Dr. Marie Tarrant, RN MPH PhD

Corresponding Author's Institution: University of Hong Kong

First Author: Marie Tarrant, RN MPH PhD

Order of Authors: Marie Tarrant, RN MPH PhD; James Ware, MA MB FRCS DMSc

Manuscript Region of Origin: HONG KONG

Abstract: In multiple-choice tests, four-option items are the standard in nursing education. There are few evidence-based reasons, however, for MCQs to have four or more options as studies have shown that three-option items perform equally as well and the additional options most often do not improve test reliability and validity. The aim of this study was to examine and compare the psychometric properties of four-option items with the same items rewritten as three-option items. Using item analysis data to eliminate the distractor with the lowest response rate, we compared three- and four-option versions of 41 multiple-choice items administered to two student cohorts over two subsequent academic years. Removing the non-functioning distractor resulted in minimal changes in item difficulty and discrimination. Three-option items contained more functioning distractors despite having fewer distractors overall. Existing distractors became more discriminating when infrequently selected distractors were removed from items. Overall, three option-items perform equally as well as four-option items. Since three option-items require less time to develop and administer and additional options provide no psychometric advantage, teachers are encouraged to adopt three-option items as the standard on multiple-choice tests.

A COMPARISON OF THE PSYCHOMETRIC PROPERTIES OF THREE- AND FOUR-OPTION MULTIPLE-CHOICE QUESTIONS IN NURSING ASSESSMENTS

Marie Tarrant RN MPH PhD[1]
Assistant Professor
Department of Nursing Studies
4/F, William M. W. Mong Block
Li Ka Shing Faculty of Medicine
21 Sassoon Road, Hong Kong
Tel: +852 2819 2643
Fax: +852 2872 6079
Email: tarrantm@hku.hk

James Ware MA MB FRCS DMSc
Director of Medical Education
Centre of Medical Education, Faculty of Medicine
Health Sciences Centre, Kuwait University
PO Box 24923
Safat, 13110 Kuwait

[1]Corresponding author

Word Count: 3146 words

Short Title: Three-option multiple-choice questions

Keywords: multiple-choice questions, multiple-choice tests, distractors, item analysis, test construction, number of choices per item, item discrimination, assessment.

Acknowledgements:
Financial support for this study was provided by the Leung Kau Kui / Run Run Shaw Research and Teaching Endowment Fund, the University of Hong Kong.

Introduction

In nursing education, multiple-choice questions (MCQs) are one of the most popular written assessment formats. Single best-answer MCQs consist of a question, two or more choices from which examinees must choose the correct option (the distractors), and one correct or best response. While MCQs are often criticized for largely assessing factual recall over higher cognitive thinking (Pamplett and Farnhill, 1995), MCQs still offer many advantages when compared with other types of written assessment. Despite what many teachers believe, MCQs are adaptable to different, although not all, levels of learning outcomes (Gronlund and Waugh, 2008). High quality MCQs present clinical vignettes to students that mimic actual clinical problems and assess application of knowledge rather than simple factual recall (Case and Swanson, 2003). Therefore, well constructed MCQs can accurately discriminate between high- and low-ability students (Schuwirth and van der Vleuten, 2003). MCQs are objective and they allow teachers to test a wider range of content and educational objectives than many other written assessment methods. Additionally, MCQs allow teachers to efficiently assess large numbers of candidates as they are easy to administer and score (McCoubrie, 2004). Furthermore, because of this broader sampling of content and because MCQ test items can be subjected to post-test review using item analysis procedures, MCQ tests have higher validity than other test methods such as short-answer or essay-style questions (Gronlund and Waugh, 2008).

Four-option MCQs remain the standard in nursing, both on in-house developed tests (Tarrant et al., 2006) and in test banks and text books used in nursing education (Masters et al., 2001). In other health science disciplines, such as

medicine, five-option items are more common (Haladyna and Downing, 1993). Although measurement specialists have long discovered that there are few evidence-based reasons for MCQs to have four or five options, many introductory books on item writing continue to recommend this practice and a majority of teachers continue to follow this recommendation (Owen and Froman, 1987). Three-options items however, have many advantages over four- and five-option items, including less time required to construct items and less testing time required. Alternately, with less time required to complete three-option items, teachers are able to increase the number of items administered on a test and thereby increase the amount of content tested (Haladyna and Downing, 1993). Furthermore, researchers have shown that in both teacher-generated (Tarrant et al., 2009) and professionally-developed (Haladyna and Downing, 1993) four-and five-option MCQs, students rarely select more than two or three of the options.

In most nursing programmes, the amount of content that requires assessment can be overwhelming. A substantial proportion of a teacher's time is spent on developing written assessments and since a substantial proportion of those assessments will likely contain MCQs, it is important that teachers are basing those practices on the best available research evidence. Additionally, student numbers are generally increasing to meet workplace shortages, while at the same time the number of available teaching faculty is often getting smaller (Broome, 2009). Therefore, because of their efficiency and ability to assess different learning outcomes, MCQs are likely to continue to remain an important component of written assessment in many nursing programmes for the foreseeable future. Thus if the time required to develop multiple-choice tests can be reduced without reducing the

reliability and validity of the assessment, this is an important consideration for nursing faculty.

Background

Numerous research studies have compared three-, four-, and five-option MCQs and most have found that three-option items perform equally as well or better than either four- or five-option items. Sidick, Barrett, & Doverspike (1994) rewrote 68 five-option items on public sector employment tests by removing the two least functional distractors. Overall, there was little difference in the psychometric properties between the three- and five-option items. Similarly, Rogers and Harley (1999) rewrote 31 four-option items on a senior secondary school mathematics test into three-option items by eliminating the least functioning distractor. The test with three-option items was less difficult than but equally as discriminating and reliable as the four-option test. As part of pre-college admissions testing, Trevisan et al. (1991) administered the same 45 items in three-, four-, and five-option formats. The three-option test form contained more highly discriminating items and fewer items with non-functioning distractors than the four- or five-option test forms. Owen and Froman (1987) randomly administered 100 items to 114 undergraduate psychology students as either five-option items or three-option items and found no significant differences in either item discrimination or difficulty. Crehan et al. (1993) eliminated the least functional distractor in 12 four-option items and found no differences in discrimination between the three- and four-option formats, although again three-option items were slightly less difficult. In one of the few studies done in a health-science discipline, Cizek and O'Day (1994) reduced 31 five-option items to four-option items on a medical specialty examination by removing a non-functioning

distractor. Study findings were consistent with other research in that four-option items were less difficult and equally as discriminating and reliable as five-option items. A recent meta-analysis (Rodriguez, 2005) and a review of research (Haladyna et al., 2002) on the optimal number of options in MCQs both concluded that in most educational settings, three-option items perform best.

By using item-analysis data and eliminating non-functioning distractors from MCQs, four or five-option items can easily be reduced to three-option items. Rogers and Harley (1999) have called for additional studies that examine the impact of reducing the number of distracters on item psychometric properties. To date, no such studies have been conducted in nursing and only one has been conducted in medicine (Cizek and O'Day, 1994). Therefore, the purpose of this study was to examine and compare the psychometric properties of four-option items with the exact same items rewritten as three-option items in nursing assessments.

Methods

Data for this study consisted of two tests administered to two cohorts of students in an undergraduate public health nursing course over two subsequent academic years. The first test consisted of 50 four-option items administered to 36 students at the end of the fall semester in 2006. The second test consisted of 70 three-option items administered to a subsequent cohort of 106 students at the same time the next year. Using item analysis data from the four-option test administered in 2006, the first author reduced the number of options to three by eliminating the least frequently selected distractor. A subset of 41 items was used on both tests. Items for both tests were identical except for the removed option and the course teacher and course content were the same for both 2006 and 2007. Tests were

criterion-referenced and absolute passing scores (50%) were used. Ample time was given to complete both tests (three hours) and all students completed the tests within the allotted time.

Item analysis data from both tests was generated using Ideal 4.1, an item-analysis software program (Precht et al., 2003) and then imported into Stata 9.2 (StataCorp, 2005) for data analysis. The Ideal program generates item difficulty and discrimination statistics, distractor performance statistics, test reliability coefficients, and mean test scores, along with other item and test performance indicators. Item difficulty is the proportion of examinees answering the item correctly, with lower values reflecting more difficult items. Items of moderate difficulty (.40 to .80) are preferable (Osterlind, 1998). Item discrimination is a measure of how effectively an item discriminates between high- and low-ability examinees (Haladyna, 2004). Discrimination is computed using either the point-biserial correlation coefficient, the correlation between the item and total test score (Osterlind, 1998) or the more simple item discrimination index, the difference in the proportion of responses between the upper and lower 27% of examinees (Ebel and Frisbie, 1991). Both the point-biserial and the discrimination index are highly correlated and discrepancies between the two statistics are extremely small or nonexistent (Beuchert and Mendoza, 1979, Oosterhof, 1976). The discrimination index was used in this analysis because it is simple to compute and explain (Ebel and Frisbie, 1991). Items are considered discriminating if the discrimination index for the correct response is positive and the same statistic for the distractors is negative. Items with higher discrimination are more desired, although recommendations for acceptable indices vary. The following categories were used to classify the item discrimination in this

study: <.10 poor; .10 to .19 low; .20 to .29 acceptable; .30 to .39 good; and ≥.40 excellent (Ebel and Frisbie, 1991, Trevisan et al., 1991). We evaluated distractor performance using two criteria to define non-functioning distractors: those chosen by fewer than 5% of examinees and those with a positive discrimination index (Rodriguez, 2005).

The psychometric properties of both tests and the subsets of 41 items were compared using descriptive statistics. We compared the mean item difficulty and discrimination of the 41 items on the two tests using the paired t-test and product moment correlations (Pearson's *r*). We also compared item difficulty and discrimination using the previously defined categories with chi-square statistics. For both tests, we evaluated distractor performance by assessing the following distractor characteristics: the proportion of distractors with low selection frequency (<5%) and positive discrimination (≥0); the proportion of functioning distractors per test; the proportion of items with 0, 1, 2, and 3 functioning distractors; and the mean number of functioning distractors per item. Finally, we evaluated the effect of removing the least frequently selected option by comparing individual distractor performance on both tests using chi-square statistics.

The unit of analysis for this study was the test item and no identifying participant information was used in any part of the analysis. Since the Institutional Review Board of the participating institution approves only human subjects' research, this study was exempted from the ethical review process. During the 2007 administration of the test, however, students were given a choice of having either the traditional test with 4-option items or a test with three-option items. All students

preferred the three-option test and students were informed that results from the test comparison could potentially be used for future publication.

Results

Table 1 shows the summary characteristics of the two tests and the subsets of 41 items. In total, 142 students were tested. On the original tests, overall mean test scores and the range of test scores were similar for both the 2006 and 2007 cohorts. The pass rate for the 2007 cohort was marginally lower than the 2006 cohort and the reliability was lower for both subsets of 41 items when compared with the whole test. However, this would be expected with fewer test items. The 41-item subset of three-option items, however, was more reliable than the subset of four-option items (.71 vs. .65).

Mean item difficulty values indicate that overall, the 41 three-option items were more difficult than the four-option items (.70 ± .15 vs. .73 ± .14) but the difference was not statistically significant (t=1.95; p=.06) (data not shown). Figure 1 presents a categorical comparison of item difficulty between the two 41-item subsets delivered in 2006 and 2007. Overall, the three-option test contained a greater number of items of moderate difficulty and fewer easy items. However, item difficulty on the two tests was similar and again this difference was not statistically significant. Differences in item discrimination present a similar picture. Mean values show that three-option items were marginally more discriminating than four-option items (.26 ± .13 vs. .25 ± .14), although again the difference was not statistically significant (t=-0.76; p=.45) (data not shown). When examined categorically, there was no significant difference in the discrimination index of three- and four-option items (Figure 2). Pearson's correlations between item difficulty and item

discrimination on both item subsets were r=.82 (p<.001) and r=.51 (p<.001) respectively.

Distractor performance is highlighted in Table 2. A substantially higher proportion of items on the three-option test were classified as functioning when compared with the four-option test (74.4% vs. 21.1%). Similarly, 56.1% of items on the three-option test had two functioning distractors compared with only 36.6% (having two or more) on the four-option test. Despite having fewer distractors, three-option items had more functioning distractors per item than four-option items (1.49 ± .64 vs. 1.32 ± .85).

Changes in distractor performance in three-and four-option items are presented in Tables 3 and 4. The removal of distractors with the lowest response frequency from the four-option items had little impact on the response frequencies of the same distractors in the three-option items. Options that were infrequently selected (<5%) on four-option items were similarly as likely to be infrequently selected on three-option tests (14.6% vs. 17.1%; p=.76) and (17.1% vs. 22.0%; p=.58) (Table 3). Reducing the number of distractors, however, did have a substantial impact upon distractor discrimination. A greater proportion of distractors were poor discriminators in four-option items when compared with three-option items (34.2% vs. 14.6%; p=.04) and (34.2% vs. 17.1%; p=.08) (Table 4).

Discussion

To our knowledge, this is the first study in nursing and only the second in a health-science discipline (Cizek and O'Day, 1994) to specifically compare item characteristics of three- and four-option MCQs. Although findings from this study are consistent with other research on this topic, generalizability may be limited by

several factors. Since our study examined only two undergraduate nursing examinations, further research in other settings should be done to determine the applicability of our findings. Also, because the number of examinees taking both tests was uneven with substantially fewer taking the four-option test, this may have affected the selection of options that were eliminated from the four-option test. Additionally, because we did not control for examinee ability it is possible that differences in the abilities between the two student cohorts may have accounted for some of the findings of this study.

The results of this study, however, do add to the growing body of research supporting three-option items. Overall, the differences in item difficulty and discrimination between four-option items and the same items rewritten as three-option items were small and statistically non-significant. Non-significant results, however, are just as important as significant results. The finding that three-option items perform equally as well as four-option items can have substantial impact upon the practice of item-writing. While there are minimal psychometric differences in item performance characteristics, clearly, three-option items are more efficient to write and administer. Aamodt & McShane (1992) estimate that students can complete an additional 12.4 three-option MCQs in the same time required to complete 100 four-option items. More items also increases test reliability. Furthermore, generating three or four plausible distracters per item is time consuming and if each distractor takes five minutes to generate, writing only three-option items would save over 16 hours of time on a 100-item test (Aamodt and McShane, 1992). Studies have found that students (Owen and Froman, 1987) and teachers (Rogers and Harley, 1999) overwhelmingly prefer items with fewer options.

Given the strong empirical and theoretical support for three-option items, Owen and Froman (1987) advise test writers to stop struggling to invent fourth or fifth options when three is almost always sufficient. Furthermore, item analysis data, if available, can be used to effectively eliminate non-functioning distractors from existing MCQs so that testing time can be reduced or content sampling can be increased.

Although reducing the number of options had minimal impact on item performance, there were positive effects on distractor performance. First, the proportions of distractors with low selection frequencies and poor discrimination were lower for three-option items. Second, fewer three-option items had 0 or 1 functioning distractors. Third, even though the total number of distractors per item was fewer, three-option items had a greater mean number of functioning distractors per item (1.49 vs. 1.32). Finally, existing distractors became more discriminating when infrequently selected distractors were removed from items. These findings illustrate that there is little benefit of including non-functioning distractors in multiple-choice items. In item writing it is challenging to come up with three or more plausible distractors to the correct answer. Consequently, item writers often add superfluous distractors that are so implausible they are selected by only a small proportion of examinees. This study has demonstrated, and others have pointed out, that a three-option item with two functioning distractors is clearly preferable to a four-option item with two or three non-functioning distractors (Schuwirth and van der Vleuten, 2004). Poorly written and clearly implausible distractors may also unintentionally cue test-wise examinees to the correct answer (Owen and Froman, 1987). Consequently, implausible distractors can introduce construct-irrelevant variance (CIV) into the assessment of student outcomes. CIV is the introduction of

extraneous variables, such as clueing or testwiseness, that are irrelevant to the construct being measured (Downing, 2002) but which can significantly affect examinee test scores (Tarrant and Ware, 2008).

Despite years of research supporting fewer options in multiple-choice items, nurse educators have not adopted the shorter items and four-option items remain the norm. Why three-option items have not been widely adopted when they are easier to write and as psychometrically robust as items with more options, is unclear. One possible explanation is that few nurse academics in the health professions have higher education in educational methods such as item construction and are therefore unaware of the literature supporting three-option items. Intuitively, three-option items would appear to be easier for examinees and thus significantly inflate student grades and pass rates. The effect of guessing on multiple-choice tests scores, however, is often overestimated (Rodriguez, 2005). Several studies have shown that reducing items from four or five to three options resulted in a test-score increase of only 1–1.2% (Aamodt and McShane, 1992, Tarrant et al., 2009). If academics are unaware of the research refuting these assumptions about guessing, however, they are unlikely to adopt three-option items. Additionally, policies regarding test format and the number of options in MCQs may not be set by the teacher but by the institutional administrators, who for the same reasons identified above, are reluctant to use fewer than four or five options on summative tests. Finally, the focus of nursing education has traditionally been more on the "what" than the "how." Although, there is an increasing focus on innovative educational methods and strategies, this is a more recent phenomenon. As universities and academics

increasingly look to evidence-based methods to deliver educational programs, item-writing practices may also become more evidence-based.

Conclusion

Results from this study of teacher-generated MCQs lends further support to the conclusion that in most circumstances, three-option items are the more feasible and practical choice when compared with four-option items. Given the time constraints of most nursing faculty today, and the increasing focus on evidence-based education, teachers involved in developing MCQs for nursing assessments are encouraged to use three-option items. Three-option items perform equally as well as the longer four-option items, they require less time to write, and the performance of the remaining distractors improves when implausible options are removed. Time spent writing four and five options is not time well spent and could be used to develop more items rather than more options. Writing tests with more items would increase the amount of content covered in the test, improve the overall reliability and validity of the test, and thus more accurately reflect student achievement.

References

Aamodt, M.G., McShane, T., 1992. A meta-analytic investigation of the effect of various test item characteristics on test scores. Public Personnel Management 21 (2), 151-160.

Beuchert, A.K., Mendoza, J.L., 1979. A Monte Carlo comparison of ten item discrimination indices. Journal of Educational Measurement 16 (2), 109-118.

Broome, M.E., 2009. The faculty shortage in nursing: global implications. Nursing Outlook 57 (1), 1-2.

Case, S.M., Swanson, D.B., 2003. Constructing written test questions for the basic and clinical sciences. National Board of Medical Examiners, Philadelphia, PA.

Cizek, G.J., O'Day, D.M., 1994. Further investigations of nonfunctioning options in multiple-choice test items. Educational and Psychological Measurement 54 (4), 861-872.

Crehan, K.D., Haladyna, T.M., Brewer, B.W., 1993. Use of an inclusive option and the optimal number of options for multiple-choice items. Educational and Psychological Measurement 53 (1), 241-247.

Downing, S.M., 2002. Threats to the validity of locally developed multiple-choice tests in medical education: construct-irrelevant variance and construct underrepresentation. Advances in Health Sciences Education 7 (3), 235-241.

Ebel, R.L., Frisbie, D.A., 1991. Essentials of educational measurement. Prentice Hall, Englewood Cliffs, N.J.

Gronlund, N.E., Waugh, C.K., 2008. Assessment of student achievement. Pearson, Upper Saddle River, N.J.

Haladyna, T.M., 2004. Developing and validating multiple-choice test items. Lawrence Erlbaum, Mahwah, NJ.

Haladyna, T.M., Downing, S.M., 1993. How many options is enough for a multiple-choice test item? Educational and Psychological Measurement 53 (4), 999-1010.

Haladyna, T.M., Downing, S.M., Rodriguez, M.C., 2002. A review of multiple-choice item-writing guidelines for classroom assessment. Applied Measurement in Education 15 (3), 309-334.

Masters, J.C., Hulsmeyer, B.S., Pike, M.E., Leichty, K., Miller, M.T., Verst, A.L., 2001. Assessment of multiple-choice questions in selected test banks accompanying text books used in nursing education. Journal of Nursing Education 40 (1), 25-32.

McCoubrie, P., 2004. Improving the fairness of multiple-choice questions: a literature review. Medical Teacher 26 (8), 709-712.

Oosterhof, A.C., 1976. Similarity of various item discrimination indices. Journal of Educational Measurement 13 (2), 145-150.

Osterlind, S.J., 1998. Constructing test items: Multiple-choice, constructed-response, performance, and other formats. Kluwer Academic Publishers, Boston.

Owen, S.V., Froman, R.D., 1987. What's wrong with three-option multiple choice items? Educational and Psychological Measurement 47 (2), 513-522.

Pamplett, R., Farnhill, D., 1995. Effect of anxiety on performance in multiple-choice examinations. Medical Education 29, 298-302.

Precht, D., Hazlett, C., Yip, S., Nicholls, J., 2003. International Database for Enhanced Assessments and Learning (IDEAL-HK): Item analysis users' guide. IDEAL-HK, Hong Kong.

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.

Rogers, W.T., Harley, D., 1999. An empirical comparison of three- and four-choice items and tests: susceptibility to testwiseness and internal consistency reliability. Educational and Psychological Measurement 59 (2), 234-247.

Schuwirth, L.W.T., van der Vleuten, C.P.M., 2004. Different written assessment methods: what can be said about their strengths and weaknesses? Medical Education 38 (9), 974-979.

Sidick, J.T., Barrett, G.V., Doverspike, D., 1994. Three-alternative multiple choice tests: An attractive option. Personnel Psychology 47 (4), 829-835.

StataCorp, 2005. Stata Statistical Software: Release 9.2. StataCorp LP, College Station, Tx.

Tarrant, M., Knierim, A., Hayes, S.K., Ware, J., 2006. The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. Nurse Education Today 26 (8), 662-671.

Tarrant, M., Ware, J., 2008. Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. Medical Education 42 (2), 198-206.

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.

Trevisan, M.S., Sax, G., Michael, W.B., 1991. The effects of the number of options per item and student ability on test validity and reliability. Educational and Psychological Measurement 51 (4), 829-837.

Table(s)

Table 1 Characteristics of the tests

	2006 Test 4-option items 36 Examinees		2007 Test 3-option items 106 Examinees	
	Original Test 50 items	Subset of 41 items	Original Test 70 items	Subset of 41 items
Mean test score % (SD)	70.3 (11.61)	--	69.7 (9.84)	--
Range of test scores (%)	38.0 – 94.0	--	41.4 – 94.3	--
Pass Rate	97.2%	--	94.4%	--
KR20 Reliability	.74	.65[a]	.75	.71[a]

SD = standard deviation; KR-20 = Kuder-Richardson 20

[a] Spearman-Brown formula used to estimate reliability for length of original test

Table 2 Distractor performance

	2006 Test 41 Items 4-option items	2007 Test 41 Items 3-option items
No. of distractors	123	82
Distractors with: n (%)		
Frequency <5%	46 (37.4)	16 (19.5)
Discrimination ≥0	60 (48.8)	13 (15.9)
Both	9 (8.5%)	8 (27.6)
Functioning distractors n (%)	26 (21.1)	61 (74.4)
Functioning distractors per item n (%)		
None	6 (14.6)	3 (7.3)
One	20 (48.8)	15 (36.6)
Two	11 (26.8)	23 (56.1)
Three	4 (9.8)	--
Functioning distractors per item M (SD)	1.32 (.85)	1.49 (.64)

Table 3 Cross-tabulation of choice frequency of distractors in three-option and four-option items

	Distractor 1		Distractor 2	
	4-option items	3-option items	4-option items	3-option items
Choice Frequency				
<5%	6 (14.6)	7 (17.1)	7 (17.1)	9 (22.0)
≥5%	35 (85.4)	34 (82.9)	34 (82.9)	32 (78.0)
	$\chi^2(1, N=82) = .091$, $p = 0.76$		$\chi^2(1, N=82) = .311$, $p = 0.58$	

Table 4 Cross-tabulation of distractors discrimination in three-option and four-option items

	Distractor 1		Distractor 2	
	4-option items	3-option items	4-option items	3-option items
Distractor Discrimination				
≥0 (poor)	14 (34.2)	6 (14.6)	14 (34.2)	7 (17.1)
<0 (good)	27 (65.8)	35 (85.4)	27 (65.8)	34 (82.9)
	χ^2 (1, N=82) = 4.23, p = 0.04		χ^2 (1, N=82) = 3.14, p = 0.08	

Figure(s)

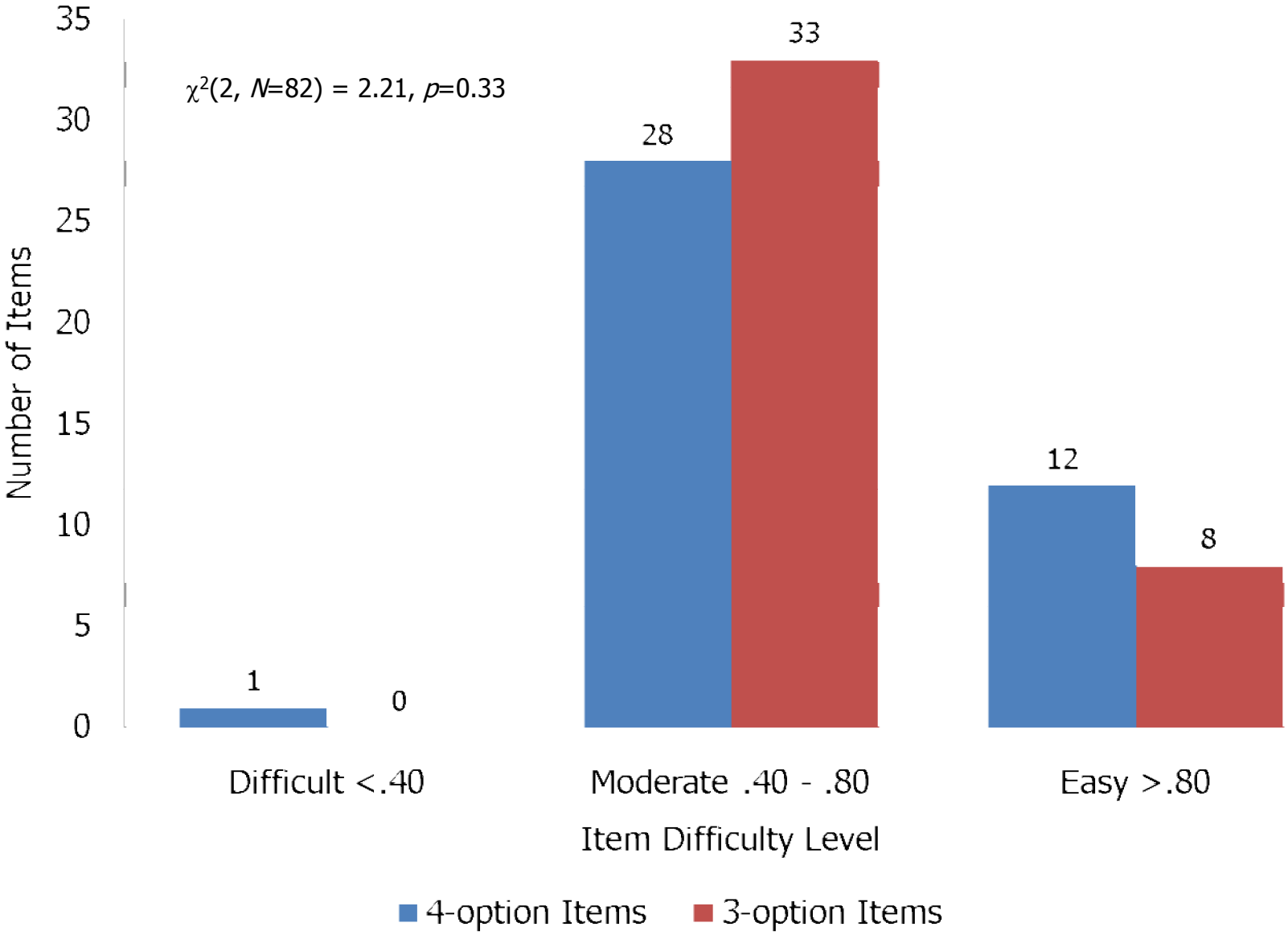

Figure 1 Comparison of item difficulty between three- and four-option items

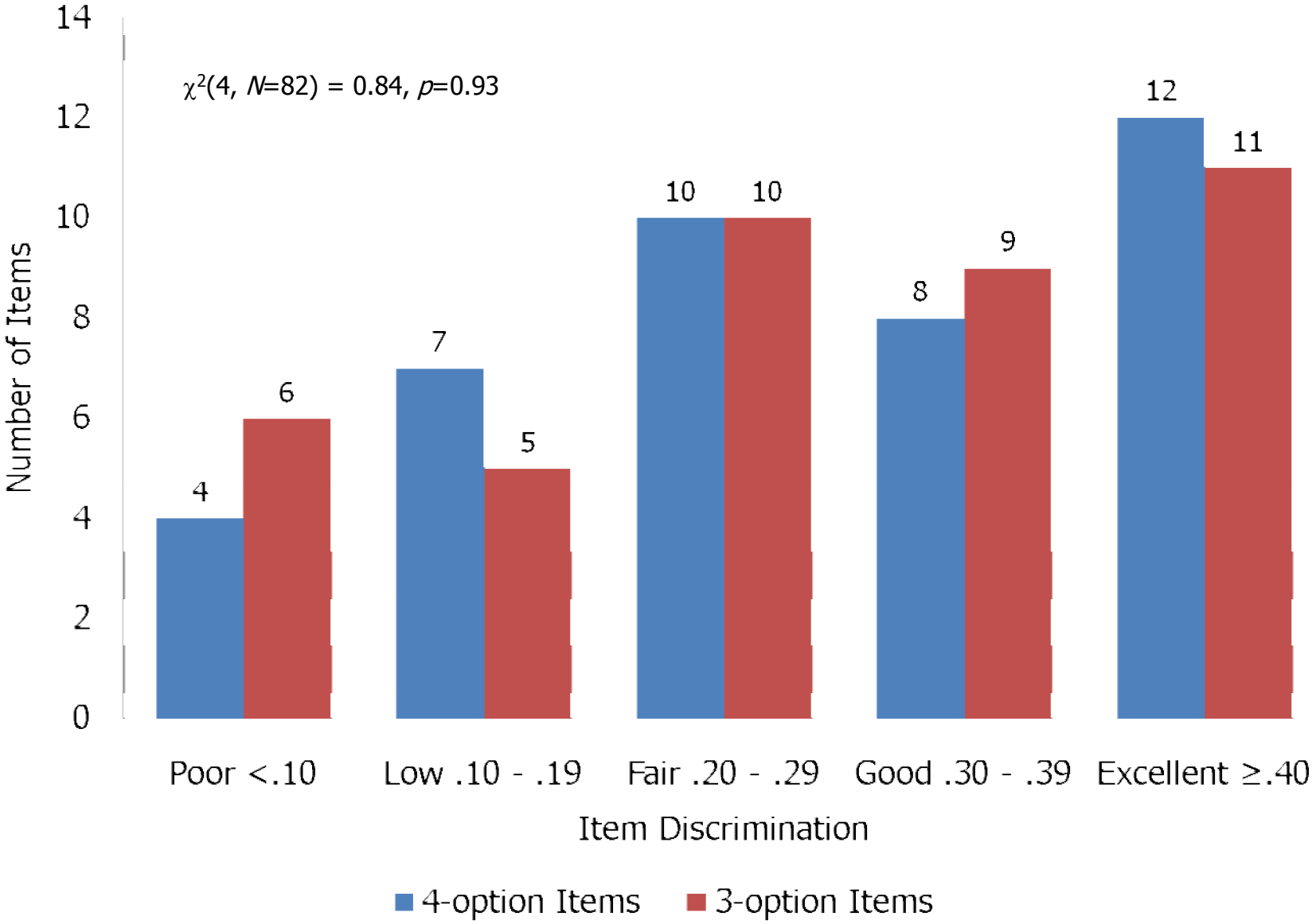

Figure 2 Comparison of item discrimination between three- and four-option items