A comparison of three- and four-option English tests for university entrance selection purposes in Japan

Tetsuhito Shizuka,
Osamu Takeuchi,
Tomoko Yashima,
Kiyomi Yoshizawa
Kansai University

The present study investigated the effects of reducing the number of options per item on psychometric characteristics of a Japanese EFL university entrance examination. A four-option multiple-choice reading test used for entrance screening at a university in Japan was later converted to a three option version by eliminating the least frequently endorsed option in each item, and was given to a separate group. Responses to the two tests indicated that using three options instead of four did not significantly change the mean item facility or the mean item discrimination. Distractor analyses revealed that whether four or three options were provided, the actual test-takers' responses spread, on the average, over about 2.6 options per item, that the mean number of functioning distractors was much lower than 2, and that reducing the least popular option had only a minimal effect on the performance of the remaining options. These results suggested that three-option items performed nearly as well as their four-option counterparts.

I. Context of the research

In Japan, each university conducts entrance screening primarily through paper-and-pencil examinations of its own creation. The results of these tests are of paramount importance since they are very often practically the only data on which admission decisions are made, except for a small number of candidates immediately above and below the cut-off line, for whom other data such as school records and extracurricular activities are also considered.

Japanese universities consider banking or reusing test items a potential threat to test security and fairness; consequently, test papers are made public immediately after administration. This results in a situation where each university develops several new test forms ever year. In the case of private universities – which traditionally admit larger numbers of students than state-run institutions – most if not all items need to be machine-scored because the time allowed for scoring all the answer sheets of one test form is often severely limited, sometimes to a day or two. Therefore, as Brown (1995) notes, selected response items –multiple-choice items in particular – are the norm.

Typically, university entrance tests are developed by in-house committees consisting of selected faculty members, who have other research, teaching, and administrative obligations as well. Writing quality multiple-choice items is a time-consuming task even for professional item writers, so, predictably, overexertion of these faculty members who have to produce new tests in full every year is a serious concern among many institutions. If it is possible to streamline the item-writing procedure without sacrificing the test quality, it would be a benefit to all concerned. The present study addressed attainability of that goal by means of reducing the number of options per item.

II. Background

Reviewing multiple-choice items used in internationally established ESL/EFL tests (e.g. TOEFL, TOEIC, Cambridge ESOL exams), one almost always finds four options per item. There is little doubt that writing four good options is considered desirable – and possible – by many authors (e.g. Gronlund, 1988; Linn and Gronlund, 2000). Haladyna *et al.* (2002) reviewed 27 textbooks on educational testing published since 1990 and showed that most advocate the idea of writing as many plausible distractors as possible. In the face of this overwhelming preference for writing four or even five options by standardized-test developers and measurement-textbook writers, it is surprising to learn that the majority of studies that have addressed this issue in the context of general educational measurement actually report results favoring three-option items.

Investigations into this matter can be classified by the approaches that the studies took. First, there were studies that tried to theoretically identify the optimal number of options. In his seminal paper, Tversky (1964) gave mathematical proof that, given a fixed total number of options for the whole test, the use of three-option items will maximize what he called the 'discrimination capacity', the 'power', and the 'uncertainty index' of a test. Discrimination capacity is defined as the number of possible distinct response patterns of a given test, power as 1 minus the probability of attaining perfect performance by chance alone, and the uncertainty index is a measure of information gained from the test. For a test consisting of k A-option items, all these indices are a function of the value A_k. When the value k x A is fixed, A_k is maximized when $A = e = 2.718$. For integer values of A, the maximum value is obtained when A=3. More recently, an information theory approach was taken by Bruno and Dirkzwager (1995), who reached the same conclusions as Tversky (1964). In an information theoretical framework, the amount of information generated when one of A options to a multiple-choice item is chosen is defined as $\log_2 A$. Hence, the mean amount of information produced *per option* is given by $(1/A) \log_2 A$. When A needs to be an integer, $(1/A) \log_2 A$ is maximized when A equals 3. Thus, Bruno and Dirkzwager (1995) showed that three-option items are the most cost-efficient in generating information. The effects of examinee ability on the optimal number of options per item were brought into light by Lord (1977). Using a 3-parameter IRT model, he took item parameters estimated for a real test and created four hypothetical tests with five, four, three, and two options per item, by replacing the guessing parameters by .20, .25, .33, or by .50, respectively. The item characteristic curves indicated that, while in the mid-ability range three-option items are the most efficient, two-option items work better at high ability levels and five-option items are superior at the lowest ability levels. In summary, irrespective of approaches taken, these theoretical studies agreed that the optimal number of options to a multiple-choice test is three, though breakdown of the group by ability may require further elaboration.

In addition to the theoretical work reviewed above, a number of empirical investigations have also been conducted in order to compare the psychometric characteristics of tests that differ in the number of options per item. Table 1 presents a very brief summary of

studies that investigated the performance of three-option items. In one study (Trevisan *et al.*, 1994), a two-option test was used to create three-, four-, and five-option versions by adding one, two, and three options respectively to each item, following item writing rules outlined by Haladyna and Downing (1989). In all the other studies, three-option items were created by discarding one (or two) of the distractors for four- (or five-) option items. The method of distractor elimination adopted can be used to classify these into the following three groups:

- studies in which the least discriminating distractors were eliminated: Williams & Ebel, 1957; Owen and Froman, 1987; Trevisan *et al.*, 1991; Crehan *et al.*, 1993;

- those in which the least frequently endorsed distractors removed: Sidick *et al.*, 1994; Delgado and Prieto, 1998; those that depended on non-empirical bases for choosing distractors to be discarded: Costin, 1970; 1972; Straton Catts, 1980; Green *et al.*, 1982; Landrum *et al.*, 1993.

Table 1 Summary of empirical studies comparing psychometric properties of items with different numbers of options per item

	Method			Results		
	Test content	Number of options per item	Distractor elimination	Item facility	Item discrimination	Test reliability
Trevisan *et al.* (1994)	Music, art, civics, etc.	5, 4, 3	N/A	5 < 4 < 3	–	5 . 4 . 3
Williams and Ebel (1957)	Vocabulary knowledge	4, 3, 2	LD	4 < 3 < 2	4 > 3 > 2	4 . 3 . 2
Owen and Froman (1987)	Psychology knowledge	5, 3	LD	5 . 3	5 . 3	–
Trevisan *et al.* (1991)	Verbal ability	5, 4, 3	LD	5 . 4 < 3	–	5 . 4 . 3
Crehan *et al.* (1993)	Psychology knowledge	4, 3	LD	4 < 3	4 . 3	–
Sidick *et al.* (1994)	Reading; writing; reasoning	5, 3	LF	5 . 3	–	5 < 3 (reading) 5 . 3 (writing) 5 > 3 (reasoning)
Delgado and Prieto (1998)	Research methodology	4, 3	LF	4 .3	4 . 3	4 . 3
Costin (1970)	Psychology knowledge	4, 3	RD	4 < 3	4 < 3	4 < 3
Costin (1972)	Psychology knowledge	4, 3	RD	4 .3	4 .3	4 > 3
Straton and Catts (1980)	Economics knowledge	4, 3, 2	RD or LP	4 < 3(RD) < 3(LP) < 2(RD)	4 > 3(LP) > 3(RD) > 2(RD)	3(LP) > 4 > 3(RD) > 2(RD)
Green *et al.* (1982)	French reading	5, 4, 3	LP	5 . 4 . 3	–	4 > 5 . 3
Landrum *et al.* (1993)	Psychology knowledge	4, 3	LP	4 < 3	–	–

Notes: LD = the least discriminating distractor eliminated; LF = the least frequently endorsed distractor eliminated; RD = distractors randomly deleted; LP = the least plausible distractor eliminated; the notations '<' and '>' indicate 'difference reported to be statistically significant and meaningful'; '.' indicates 'difference reported to be statistically non-significant or significant but practically negligible.'

Despite differences in the methods used, two trends were identifiable in the results of these studies. First, item facility tended to be slightly higher in the three-option format than in the four- or five-option format. That was the case in 7 out of the 12 studies. It is noteworthy, however, that in the two studies that removed the least frequently endorsed options, no such changes were observed. Second, item discrimination and test reliability were either not much affected by changing the number of options, or were affected in a less systematic manner. Only 3 out of 7 discrimination comparisons reported meaningfully higher discrimination for the four-option format. Regarding reliability, only 2 out of11 cases favored four or five options over three. Many of the authors of these studies concluded by defending and/or promoting the three-option format, either because they found no significantly different item performance between the three-option format and formats with more options or because, even when they did, the effect size was negligible. The following statement by Delgado and Prieto (1998: 200–01) was representative: 'From an empirical point of view, the three-option format is at least as

defensible as its four-option counterpart [and] taking into account pragmatic considerations such as the economy of time and energy, the three-option format is clearly better'.

The slight and occasional facility increase may be attributed to increased levels of chance success, but why was discrimination not very much affected? Distractor analyses conducted by Haladyna and Downing (1993) appear to provide some answer. Examining four standardized multiple-choice tests, whose number of options per item was five or four, they found that the average number of functioning distractors per item was 1.42, .92, 1.29, and .91 for the four tests, respectively. Items with three effective distractors were very rare, occurring only from 1.1% to 8.4% of the time. That is, most of the 'job' was being done by two or fewer functioning distractors.

It would be fair to say, then, that adopting the three-option format demands much more serious consideration than is commonly given by writers of standardized tests and entrance examinations in EFL circles. To the best of our knowledge, few studies have addressed this issue in the field of language testing. The present study investigated the extent to which the viability of the three-option format would hold in the context of Japanese EFL university entrance selection. Specifically, it examined the effects of reducing the number of options per item from four to three on the mean item facility, the mean item discrimination, and distractor performances.

III. Method

1. The four-option test

One test administered for entrance screening at one of the faculties of a major private university in western Japan in 2003 was used for this study. Table 2 shows the specifications for this test. From the data produced by all the applicants who took this test, responses made by 1000 were randomly sampled for this study. As can be seen in Table 2, the actual test used for entrance selection included not only four-option items designed to tap local/global reading comprehension but also sequencing items intended to tap command of syntax in production and constructed-response translation items. Since the focus of this study was on the number of options in multiple-choice items, the sequencing and translation items were not relevant. After responses to these items were removed, retained as the baseline data for this study were 38 000 data points produced by 1000 applicants responding to 38 four-option multiple-choice comprehension items. This 38-item subset of the original test will hereafter be referred to as the 'four option test'. Although the actual test questions cannot be provided, illustrative items are given in Appendix 1.

Table 2. Specifications for the original test

Material type	Targeted constructs	Item type	*k*
Part 1			
descriptive	local comprehension	four-option multiple-choice	10
c. 600 words	global comprehension	four-option multiple-choice	3
	local comprehension	L2–L1 translation	1
Part 2			
narrative	local and global comprehension	rational cloze	16
c. 700 words		four-option multiple choice	
	global comprehension	four-option multiple choice	4
Part 3			
dialogue	local comprehension	gap filling	5
8 turns each		four-option multiple choice	5
Part 4			
argumentative	local comprehension	four-element sequencing	5
c. 200 words	constructive syntax		
	sentence writing	L1–L2 translation	1

Note: k = the number of items

2. The three-option test

The 38-item four-option test was edited to produce a streamlined version consisting of 10 four-option items and 28 three-option items using the following procedure. The first 10 items (items 1–10) were retained as they appeared in the four-option test, for the purpose of linking. For each of the remaining 28 items (items 11–38), the distractor that attracted the smallest number of test-takers in the entrance screening administration was eliminated.

The frequency distribution of the least favored distractor's endorsement percentage is shown in Table 3. The maximum was 18.7 and the minimum was .3, which translates to only 3 persons out of 1000. The mean was 6.6, the median was 4.5, and the standard deviation was 4.7.

The deletion method based on endorsement frequency, as was used in group 2 empirical studies reviewed above, was adopted in this study because it was expected to have more practical future application in our context. In Japan, as mentioned above, concern over fairness and test security makes it extremely difficult to pilot entrance examination items before the actual administration. Therefore, even if deleting empirically identified non-discriminating distractors turned out to be an effective way of streamlining a multiplechoice test, it would have very limited future applications. On the other hand, if removal of empirically identified unattractive options proved to be a successful technique, it could have some implications –though somewhat indirect – for the actual item writing phase, because research indicates that intuitive detection of potentially unpopular distractors is feasible (Cizek and O'Day, 1994). Having said that, an unpopular distractor may, in many cases, be a nondiscriminating distractor as well. In fact, point-biserial correlations between endorsement/non-endorsement of these distractors and the total score were not significant for 19 out of 28 items. That is, the least popular distractors were at the same time non-discriminating ones for about 68% of the cases.

Table 3. Frequency distribution of theendorsement

percentages of the least favored distractors

Endorsement percentage	Frequency
0.0–2.0	3
2.1–4.0	9
4.1–6.0	3
6.1–8.0	3
8.1–10.0	5
10.1–12.0	2
12.1–14.0	0
14.1–16.0	1
16.1–18.0	1
18.1–20.0	1

After these least popular options were removed, the correct options and the two remaining distractors were recoded using A, B, and C. In this way, a new version consisting of 10 four-option items and 28 three-option items was created. The first 10 items were identical in the two versions; the other 28 items were the same except that the new version had one less option per item. Even though the new version was a mixture of four- and three-option items, it will be referred to as the 'three-option test' for convenience.

3. Participants

The participants were Japanese students at the university entrance level. The average applicant to a Japanese university has six years of English as a foreign language education, exclusively in the classroom. Instruction at high schools is still largely reading-focused, rather than communication-oriented. Students who took the four-option test were applicants to one of the faculties in humanities of the university and the responses by randomly sampled 1000 were analyzed for this study. No criterion-referenced data are available regarding the English proficiency of these students, but anecdotal evidence suggests that students in this faculty tend to have higher instrumental motivation toward learning the language than those in the other faculties of the university. The three-option test was given to a total of 192 first-year students in another faculty in humanities of the same university, approximately 9 weeks after the four-option test administration. They had not taken the original test in the entrance selection process. The two faculties are demographically quite similar except that the first faculty has a larger proportion of female students.

IV. Results

1. Item facility

First, to check the comparability of the two groups, descriptive statistics of the number correct scores for the 10 common items were examined. The mean for the four-option group (n=1000) was 6.69 (SD=1.78) while that for the three-option group (n 192) was 5.88 (SD=1.90). The variances were not significantly different from each other (F=.88, n.s.), and a t-test assuming equal variances revealed a significant difference between the means (t=5.69, p<.01, df=9, two-tailed). Thus, the four-option group turned out to be significantly higher in ability, which made simple comparison of the number correct scores on items 11–38 untenable; in that part, two groups different in ability responded to two versions of items that were different in the number of options.

Therefore, common item equating in the Rasch measurement framework was performed using *FACETS v.3.0* software. First, item estimates were separately computed for the four- and the three-option datasets. When the difficulty parameters of the 10 common items were cross-plotted, the slope of the best-fit line was .72, indicating the 10-item set as a whole was less than satisfactory as the anchor item set (Linacre, 2004). Items corresponding to points far away from the best-fit line were successively dropped so that the slope would get closer to 1.00. After four items were removed, the slope was .99, so the remaining six were deemed satisfactory as anchors.

At the same time as examining qualities of common items, items unique to the four- and three-option tests were screened for obvious anomalies. This identified one item (item 36) as an apparent problem both in the four- and the three-option versions; the point-biserial correlations were negative (- .13 and -.08, respectively) and the outfit mean squares were clearly higher than model expectation (1.2 and 1.5, respectively). This item required a gap in a dialogue to be filled with an utterance containing an expression which, in hindsight, even the highest ability candidates to the university were unlikely to have learned, hence possibly eliciting random guessing behaviors. Since this anomaly had to do with the correct option of this specific item, not with the number-of-options factor, it was decided to drop this item in subsequent analyses. *FACETS* was run again on the four option test data (less the four common and the one unique items), and then on the three-option test data (less the four common and the one unique items) with parameters of the six common items anchored. Parameter displacements were all smaller than .29 logits, confirming acceptable anchoring performance (Linacre, 2004).

Fit statistics from the final runs for the two unique item sets were examined using the criteria proposed by Smith *et al.* (1998). Infit mean square values that should be flagged as misfits are those larger than 1.06 for a sample of 1000 (four-option items), and 1.14 for a sample of 192 (three-option items); corresponding critical values of outfit mean squares are 1.19 and 1.43, respectively. According to these guidelines, 4 of the 27 four-option items were diagnosed as potential misfits, while none of the three-option items were. Also, the standard deviations of infit and outfit were slightly smaller for the three-option data (.07 and .10, respectively) than for the four-option data (.09 and .13). Hence, the three-option data seemed to conform to the Rasch model at least as well as, if not more exactly than, the four-option data.

Difficulty parameters obtained for items 11–38 (less item 36) when the numbers of options were three and four are summarized inTable 4. The mean difficulty (in logits) for four-option items was.02 and that for three-option items was .20. A paired *t*-test revealed that this difference was not significant (t=-1.97, p=.06, df =26, two-tailed). Item separation reliabilities (Wright and Masters,1982) – the estimated ratios of true difficulty variances to error variances – were very high: .99 for the four-option items and .96 for the three-option items. (The higher coefficient for the four-option items indicates that they were more reliably calibrated because of a much larger number of persons.) These

high item reliabilities indicate that our findings regarding item difficulties can be considered highly generalizable.

Hence, in line with previous studies that adopted popularity-based elimination (Sidick *et al.*, 1994; Delgado and Prieto, 1998), reducing the number of options did not make the items easier; in fact the mean difficulty parameter was slightly higher for the three-option versions (though not to a significant extent). Figure 1 shows a scatter plot of four- against three-option item difficulties. Data points for the six common items naturally are plotted along the identity line. It can be visually confirmed that points for the test-unique items are scattered along the identity line. It is not the case that the data points for the unique items concentrate either above the identify line (which would mean that three-option versions were generally more difficult) or below it (which would indicate that four-option items tended to be more difficult). The Pearson product-moment correlation between the difficulty parameters for the 27 unique items was .87 (p<.01), indicating that the relative difficulties of items were reasonably stable irrespective of the number of options.

Table 4. Difficulty parameters of four- and three-option versions of items 11–38 (less item 36)

	k	Mean	*SD*	Min	Max	Range
Four-option	27	.02	.93	−2.00	1.75	3.75
Three-option	27	.20	.81	−2.14	1.51	3.65

Note: *k* = the number of items

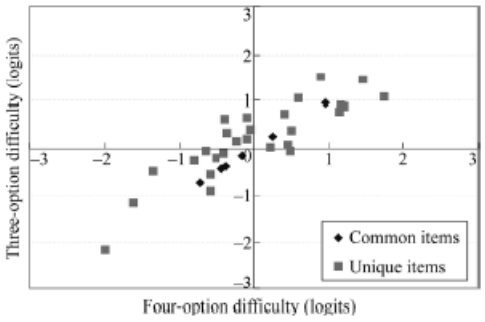

Figure 1. Plot of four-option against three-option item difficulties

2. Item discrimination

The Rasch model specifies discrimination to be uniform across all the dichotomous items in a given dataset, as a measurement ideal. However, it does not in any way presuppose that empirical discriminations of all items are equal. Examining the actual discriminations, as well as fit statistics, helps us identify where and how much the real data did or did not conform to the measurement requirements.

A straightforward but potentially misleading procedure for examining discrimination would have been to compute item-total correlations in each set and simply compare the coefficients. One problem with this method in the case of the present study was the large sample-size difference between the two sets; that for the four-option test was 1000, while that for the three-option test was 192. Item-total correlation is a function not only of the discriminatory power of the item but also of the variance of the total score, which tends to increase as the sample size gets larger. In fact, the variance of the person measures was significantly larger in the four-option group (F =1.21, p<.01), indicating that straightforward item-total computation was inappropriate.

In order to circumvent this problem, we decided to take sub-samples (n=192, each) of the larger four-option data and compare each of them with the smaller three-option data (n=192). These sub-samples were taken such that a four-option sub-sample and the three-option sample consisted of test-takers with comparable ability distributions. This requirement was met by matching the test-takers as closely as possible based on their Rasch ability estimates on the common scale (these ability estimates served as the 'total score' in the subsequent computation of item-total correlations). This procedure ensured that differences between the four- and the three-option data, if any, could be attributed to the effects of the number of options, because the ability distributions were controlled for. Three sub-samples satisfying the above-mentioned requirements were randomly extracted from the four-option datasets.

Person separation reliabilities (Wright and Masters, 1982) – the Rasch equivalents of Cronbach's alpha coefficients – of the three four-optionsubsets were all .68, and that for the three-option set was .68 as well. Two observations are possible regarding these coefficients. One is that they were rather low, necessitating cautious interpretation of the discrimination analyses below, in which item responses are correlated against the ability estimates. The other is that reducing the number of options did not affect reliability adversely. Point-biserial correlations computed for items 11–38 (less item 36) are summarized in Table 5. The means for the four-option subsets were .30–.31, while that for the three-option set was .29. One-way analysis of variance revealed that there were no significant differences among the four means. Hence, there was no evidence that the streamlining decreased item discrimination.

Table 5. Mean item discriminations of four- and three-option items

	Four-option			Three-option
	Subset A	Subset B	Subset C	
k	27	27	27	27
Mean*	.30	.31	.31	.29
*SD**	.11	.13	.13	.12
Max	.46	.54	.51	.51
Min	.03	.08	.04	.06

Notes: k = number of items; *computation performed on Fisher-*z* transformed values

3. Distractor performance

a Actual equivalent number of options

We began our distractor analyses by computing the actual equivalent number of options (AENOs) defined by Sato and Morimoto (1976). An AENO is an index of the distribution of actual option endorsements, which is given by:

$$\text{AENO} = 2^{-\sum_{i=1}^{k} P_i \log_2 P_i}$$

where P_i denotes the endorsement percentage of the *i*-th option of a *k*-option item. The value will be 1.00 when all endorsements fall on one option, 4.00 when each alternative of a four-option item attracts 25% of test-takers, and 3.00 when each option of a three-option item is endorsed by 33.3% of test-takers. Although an AENO does not necessarily reflect item quality, it does indicate the relative plausibility of options in an item. Of interest was whether or not removal of the least favored options resulted in a substantial decrease in AENO values. If not, that would be another piece of evidence that the fourth option could be dispensed with.

Table 6. Percentage distribution of AENOs for four- and three-option items

AENO	1.00–1.50	1.51–2.00	2.01–2.50	2.51–3.00	3.01–3.50	3.51–4.00
Four-option	3.70	7.41	33.33	25.93	18.52	11.11
Three-option	3.70	3.70	18.52	74.07	n/a	n/a

Percentage distributions of AENOs when the numbers of options were four and three are shown in Table 6. In the case of four-option items, where the possible maximum AENO was 4.00, just about 30% of the items had values higher than 3.00, and those with values higher than 3.50 were only 11%. On the other hand, when the number of options was three, as many as 74% of the items had AENOs between 2.51 and 3.00, the band closest to the possible maximum. The means were very similar, at 2.67 for the four-option items and 2.61 for the three-option items, whereas the standard deviation was much larger for the four-option items (*SD*=.64) than for the three-option items (*SD*=.35). In 13 out of the 27 pairs, the AENO value was larger in the three-option items. A non-parametric Wilcoxon's signed-rank test was performed, which found no significant difference between the AENOs of four- and three-option items (*z*=.45, n.s.).

b Endorsement rankings:

Next we examined whether and to what extent removing the least popular distractor from an item changed the endorsement rankings of the other three options of that item. The results showed that, in 25 out of the 27 items, the most frequently endorsed options were identical in both tests and that, in 21 of them, the rankings of all three options were identical. Option-by-option comparison revealed that in the total of 81 (=3X27) options,

68 retained exactly the same position in both tests, 12 moved up or down by one place, and only one moved up two places. That is, 84% of the options retained exactly the same relative positions and all the other options but one changed position only by one place.

c Distractor discrimination:
As stated above, AENOs by and of themselves do not reveal much about qualities of the options. What is important is not only how many but also what level of the test-takers chose which options. In the present study, when the endorsement/non-endorsement of a distractor had a significant (p<.05) negative correlation with the ability estimate, the distractor was classified as discriminating, and when otherwise, as non-discriminating. Table 7 shows the counts of items with three, two, and one discriminating distractors in each set. The possible maximum number was three for four-option subsets and two for the three-option set. One observation is that items with three discriminating distractors were quite rare in any of the four-optionsubsets. There were only one to three such items out of 27, which translates to 4–11%, corroborating Haladyna and Downing's (1993) report. Another is that items with one or two discriminating distractors occurred in the three-option set with very similar frequencies to those observed for the four-option subsets, even though the numbers of provided distractors were different. There were 25, 21, and 20 such items in the four-option subsets, and 26 in the three-option set. The four (sub) sets were comparable in the indices of central tendency as well. The mean numbers of discriminating distractors per item for the four-optionsubsets A–C and the three-option set were 1.56, 1.26, 1.44, and 1.37, respectively; the medians were 2, 1, 1, and 1, respectively. Sign tests were performed to check for statistical significance in the medians on the three pairs of datasets: four-optionsubset A vs. three-option set, four-optionsubset B vs. three-option set, and four-optionsubset C vs. three-option set. The results revealed no significant differences in the numbers of functioning distractors per item attributable to the number-of-options factor.

Table 7. The counts of items with 3, 2, and 1 discriminating distractors

Item with …	Four-option			Three-option
	Subset A	Subset B	Subset C	
3 discriminating distractors	1	2	3	n/a
2 discriminating distractors	14	7	10	11
1 discriminating distractor	11	14	10	15
0 discriminating distractors	1	4	4	1
Total	27	27	27	27

d Change in distractor discrimination:
Finally, the effect of removing the least popular distractor on the discriminations of the remaining distractors was investigated. Focusing on 54 distractors common to both tests, discriminating (p<.05) and non-discriminating (n.s.) ones were cross-tabulated. Table 8 shows the results. One observation is that, in the vast majority of cases, when a

distractor was discriminating in a four-option item, it was also discriminating in a three-option item as well. This pattern occurred in 81%, 74%, and 82% of the cases in subsets A, B, and C, respectively. Another observation is that, on the contrary, when a distractor was not discriminating in a four-option item, it was changed to a discriminating one in more than half of the cases (55%, 63%, and 54% in subset A, B, and C, respectively). Third, when cases in which the status did not change were aggregated to be compared with cases in which the status changed (for better or for worse), the former accounted for 67%, 56%, and 65% insubsets A, B, and C, respectively. To summarize, then, removing the least favored option did not affect the performance of the remaining distractors in the majority of cases; discriminating and non-discriminating ones remained discriminating and non-discriminating, respectively. When it did affect the performance, there were more cases where the non-discriminating one became discriminating than the other way around. McNemar tests of symmetry were performed to examine the statistical significance of this trend. *P*-values were .23, .06, and .06 for three-option vs. subset A, three-option vs. subset B, and three-option vs. subset C, respectively. Hence, there was no statistically significant change in discrimination/non-discrimination status, but at the same time, the *p*-values for the tests involvingsubsets B and C suggest a trend for the status to change – when it did – more often for the better than for the worse.

Table 8. Cross-tabulation of discrimination status of distractors in four- and three—option Items

	Four-option Subset A		Four-option Subset B		Four-option Subset C	
	$p < .05$	n.s.	$p < .05$	n.s.	$p < .05$	n.s.
Three-option $p < .05$	25	12	20	17	23	14
Three-option n.s.	6	11	7	10	5	12

V. Discussion

Prior to discussion of the results, limitations of the present study need to be pointed out. First, the four-option test had the translation and sequencing items interspersed, while the three-option test consisted only of multiple-choice items. This was a potential threat to the internal validity of the study. Second, it should be emphasized that the three-option test in this study was created based on four-option item statistics from the previous administration. It is not yet clear to what extent the findings are applicable to a situation where such statistics are not available. Third, the relatively low person separation reliabilities indicate that we should be careful in generalizing the discrimination-related findings. Within these limitations, what emerged from the present study were the following results.

1. Item facility

Item facilities remained practically the same even after the least popular distractors were discarded. The three-option items even appeared slightly more difficult in our data. It should be noted that the present study obtained this result by controlling variables in a

stricter manner than previous studies by adopting Rasch-based anchoring. This may at first seem counterintuitive when common belief holds that fewer options will translate to higher probabilities of chance success and, hence, higher mean scores. However, a more reasonable assumption may be that, as Downing (1992) pointed out citing Ebel's (1968) study, motivated examinees rarely resort to random guessing when they have sufficient time and the item difficulty level is appropriate. If the proportion of blind guessers is relatively small, decreasing the number of options will only have a minor impact on the mean score. The adequacy of this assumption was implied in the rather low median endorsement percentage (4.45%) of the least popular distractors in the entrance selection administration. Test-takers who end up on the distractors that very few of the others do are most likely to be low performers who choose them either because of random guessing or very serious misunderstanding. When faced with streamlined items that lack the least popular option, test-takers who normally choose those options in the four-option format are likely either to omit the item or resort to random guessing among the remaining three options (Ramos and Stern, 1973). In either case, when the number of such test-takers is small, their behavior may well only slightly affect the item facilities.

2. Item discrimination and distractor performance

With regard to item discrimination, streamlining lowered the values only minimally. This supports Delgado and Prieto's (1998) and Crehan *et al.*'s (1993) findings, and adds still another piece of evidence for the viability of three-option items created by eliminating the least frequently endorsed options. Close examination of distractor performances shed some new light on why this method of distractor removal did not much affect item discriminations or test reliability. First, the endorsement percentages of the least popular distractors were quite low. The median percentage of 4.45 means that half of these distractors were chosen by less than 5% of the examinees. This was also reflected in the lack of significant difference between the mean AENO of the four-option items and that of the three-option items. Even when presented with four options, test-takers' actual choices spread out over practically the same range – over a little more than two and a half options – as when given three options. Second, the number of functioning distractors did not change significantly whether the number of provided distractors was three or two. In the present study, the mean numbers of discriminating distractors per item were 1.26–1.56 in the four-option test and 1.37 in the three-option test. Items with three effective distractors accounted for only 4–11%. These values are quite similar to those reported by Haladyna and Downing (1993), although their definition of 'non-functioning' distractor was based on trace-line analyses, not on distractor-total correlations like ours. Third, in the vast majority of cases, discarding the least popular distractors did not affect the within-item endorsement rankings of the remaining options at all, and even when it did, it did so only very slightly. The three most popular options in each item almost always retained their respective positions whether or not accompanied by the fourth and least popular option. Fourth, the significance of distractor discrimination survived the streamlining in most cases. Discriminating distractors in the four-option format usually did discriminate in the three-option format as well. On the

other hand, when a distractor did not function in the four-option format, removing the least favored option improved the situation more often than not.

3. Advantages of the three-option format

The primary impetus behind the present study was the desire to lighten the workload of multiple-choice item writers, and its results indicate that streamlined three-option versions are likely to function equally well. If three-option items can legitimately replace their four-option counterparts, it will lead to a substantial workload reduction. If it takes five minutes to write and document each additional alternative (Sidick *et al.*, 1994), replacing the four-option format with the three-option one would save a total of 32 hours of item writers' time spent in producing ten forms of 38-item tests. Since the fourth – often implausible – option typically takes more time to come up with than the other three, the amount of time saved may be even greater.

Adoption of the three-option format is also relevant to those who are interested in increasing measurement accuracy rather than in saving item writers' time. Even if Tversky's (1964) assumption that the duration of test-taking time is proportional to the total number of options in the test does not quite hold, there is ample evidence that it takes less time to respond to a three-option item than to a four-option equivalent. Reviewing eight empirical studies with 14 samples, Aamodt and McShane (1992) have found that on average 112.4 three-option items can be completed in the same amount of time as100 four-option items can. Other things being equal, giving more items in the same amount of time should result in higher test score reliability.

In addition to item writing efficiency or enhanced test reliability, advantages of the three-option format cited in the literature (Straton and Catts 1980; Budescu and Nevo, 1985; Owen and Froman, 1987; Haladyna and Downing 1993; Delgado and Prieto, 1998; Rogers and Harley, 1999) are:

- the length of a test booklet is smaller;
- printing costs are reduced;
- the distractors taken as a set should be more plausible;
- students can answer questions with less distractions;
- students will feel less pressure because they can work more slowly or spend time to recheck; and
- the chances of providing unintended cues that profit test-wise students will be decreased.

VI. Conclusions and future directions

In summary, the results of the present study indicate that whether we provide four or three options:

- the mean item facility does not change significantly;

- the mean item discrimination as well as test reliability does not decrease significantly;
- examinees' actual choices spread out only over about 2.6–2.7 options and relative popularities of the options in each item stay mostly the same; and
- the average number of functioning distractors per item is less than two, and items with three functioning distractors are only exceptional.

The four options may be there, but are often not doing their jobs. In the light of these results, it is hard not to agree with Haladyna's (1997; 1999) suggestion that, acknowledging that three options are about as many as they can write in order for each of them to be functional, item writers should write fewer distractors and concentrate on their quality instead of their quantity. We also concur with Owen and Froman (1987: 519) when they state, 'To those who struggle, semester after semester, inventing fourth or fifth options (that are too often implausible), we offer this advice: Stop!'.

Future research needs to examine whether and to what extent two types of three-option tests differ in terms of psychometric characteristics: one produced as a three-option test from the beginning and the other created as a streamlined version of a four-option test. Another issue worth exploring is how well item writers can intuitively predict the empirical endorsement percentage of each alternative of four-option items, since the trustworthiness of such intuition will be crucial for the first type of three-option test to be successful.

Acknowledgements
This research was financially supported by the Kansai University Research Grants: Grant-in-Aid for Joint Research, 2001. The authors are grateful to the anonymous reviewers of *Language Testing* and to Joseph Pielech for their suggestions and comments on earlier versions of this article.

VII. References

Aamodt, M.G. and **McShane, T.** 1992: A meta-analytic investigation of the effect of various test item characteristics on test scores and test completion times. *Public Personnel Management* 21, 151–60.

Brown, J.D. 1995: English language entrance examinations in Japan: myths and facts. *The Language Teacher* 19, 21–26.

Bruno, J.E. and **Dirkzwager, A.** 1995: Determining the optimal number of alternatives to a multiple-choice test item: an information theoretic perspective. *Educational and Psychological Measurement* 55, 959–66.

Budescu, D.V. and **Nevo, B.** 1985: Optimal number of options: an investigation of the assumption of proportionality. *Journal of Educational Measurement* 22, 183–96.

Cizek, G.J. and **O'Day, D.M.** 1994: Further investigation of nonfunctioning options in multiple-choice test items. *Educational and Psychological Measurement* 54, 861–72.

Costin, F. 1970: The optimal number of alternatives in multiple-choice tests: some empirical evidence for a mathematical proof. *Educational and Psychological Measurement* 30, 353–58.

Costin, F. 1972: Three-choice versus four-choice items: implications for reliability and validity of objective achievement tests. *Educational and Psychological Measurement* 32, 1035–38.

Crehan, K.D., Haladyna, T.M. and **Brewer, B.W.** 1993: Use of an inclusive option and the optimal number of options for multiple-choice items. *Educational and Psychological Measurement* 53, 241–47.
Delgado, A.R. and **Prieto, G.** 1998: Further evidence favoring three-option items in multiple-choice tests. *European Journal of Psychological Assessment* 14, 197–201.
Downing, S.M. 1992: True-false and alternative-choice formats: a review of research. *Educational Measurement: Issues and Practices* 11, 27–30.
Ebel, R.L. 1968: Blind guessing on objective achievement tests. *Journal of Educational Measurement* 5, 321–25.
Green, K., Sax, G. and **Michael, W.B.** 1982: Validity and reliability of tests having different numbers of options for students of differing levels of ability. *Educational and Psychological Measurement* 42, 239–45.
Gronlund, N. 1988: *How to construct achievement tests*. 4th Edition. Prentice-Hall.
Haladyna, T.M. 1997: *Writing test items to evaluate higher order thinking*. Allyn and Bacon.
Haladyna, T.M. 1999: *Developing and validating multiple-choice test items.* 2nd Edition. Lawrence Erlbaum.
Haladyna, T.M. and **Downing, S.M.** 1989: Validity of a taxonomy of multiplechoice item-writing rules. *Applied Measurement in Education* 2, 51–78.
Haladyna, T.M. and **Downing, S.M.** 1993: How many options is enough for a multiple-choice test items?
Educational and Psychological Measurement 53, 999–1010.
Haladyna, T.M., Downing, S.M. and **Rodriguez, M.C.** 2002: A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education* 15, 309–34.
Landrum, R.E., Cashin, J.R. and **Theis, K.S.** 1993: More evidence in favor of three-option multiple-choice tests. *Educational and Psychological Measurement* 53, 771–78.
Linacre, J.M. 2004: *A user's guide to WINSTEPS/MINISTEPS: Rasch-model computer programs*. Institute for Objective Measurement.
Linn, R.L. and **Gronlund, N.E.** 2000: *Measurement and assessment in teaching.* 8th Edition. Prentice-Hall.
Lord, F. 1977: Optimal number of choices per item: a comparison of four approaches. *Journal of Educational Measurement* 14, 33–38.
Owen, S.V. and **Froman, R.D.** 1987: What's wrong with three-option multiple choice items? *Educational and Psychological Measurement* 47, 513–22.
Ramos, R.J. and **Stern, J.** 1973: Item behavior associated with changes in the number of alternatives in multiple choice items. *Journal of Educational Measurement* 10, 305–10.
Rogers,W.T. and **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, 234–47.
Sato, T. and **Morimoto, U.** 1976: Sentaku-shi keishiki tesuto kaitou bunpu no bunseki [Analyzing endorsement distribution of selected-response items]. In Proceedings of the 4th Annual Meeting of the Behaviometric Society of Japan. Behaviometric Society of Japan.
Sidick, J.T., Barrett, G.V. and **Doverspike, D.** 1994: Three-alternative multiple- choice tests: an attractive option. *Personnel Psychology* 47, 829–35.
Smith, R.M., Schumacker, R.E. and **Bush, M.J.** 1998: Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement* 2, 66–78.
Straton, R.G. and **Catts, R.M.** 1980: A comparison of two, three, and four-choice item tests given a fixed total number of choices. *Educational and Psychological Measurement* 40, 357–65.
Trevisan, M.S., Sax, G. and **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, 829–37.
Trevisan, M.S., Sax, G. and **Michael, W.B.** 1994: Estimating the optimum number of options per item using an incremental option paradigm. *Educational and Psychological Measurement*

54, 86–91.
Tversky, A. 1964: On the optimal number of alternatives at a choice point. *Journal of Mathematical Psychology* 1, 386–91.
Williams, B.J. and **Ebel, R.L.** 1957: The effect of varying the number of alternatives per item on multiple-choice vocabulary test items. In Huddleton, E.M., editor, *The 14th Yearbook of the National Council on Measurement in Education*, Michigan State University, 63–65.
Wright, B.D. and **Masters, G.N.** 1982: *Rating scale analysis*. MESA Press.

Appendix 1 Sample items in the four-option test

Part 1 Local comprehension

(The examinee chooses the option which is the closest in meaning to the underlined part in the text.)

What the author feels is '<u>not coherent</u>' is the fact that:
A) Japanese people do not flip their given and family names
B) Western names and Japanese names are treated differently
C) Japanese people put their family names before their given names
D) people's names appear in different forms in different languages

Part 1 Global comprehension

(The examinee chooses the option that best completes a statement regarding the text.)
In paragraph three, the author
A) gives historical backgrounds to the practice described already
B) presents another reason for making the changes he calls for
C) provides an example of a problem introduced in paragraph one
D) elaborates on why he believes the changes are necessary

Part 2 Multiple-choice cloze

(The examinee chooses the word that best fills the gap created in the text.)
. . . found that drivers were four times more likely to be involved in a car ( ) while talking on the phone. Studies by the National Highway Traffic Safety Administration find . . .
A) race B) show C) crash D) sale

Part 3 Dialogue gap filling

(The examinee chooses the utterance that best fills the gap created in the dialogue.)
Ken: Hi, Mary. How's your week going?
Mary: Hi, Ken.____
Ken: What's wrong?
A) You are very welcome.
B) Couldn't be better.
C) It's not going at all.
D) Not so well, I'm afraid.