教育研究學報 1993, 第八卷, 74-78

Educational Research Journal 1993, Vol. 8, pp.74-78

Does the Ordering of Questions in a Test Affect Student Performance?

David CONIAM

The Chinese University of Hong Kong

This paper examines the question of whether the ordering of questions in a test has a bearing on students' scores on that test. Folklore suggests that it is preferable to put the easier items at the beginning of the test, as this will have a positive motivational effect on students taking the test. The results of the current study indicate that this is not the case. A questionnaire administered to students indicated that while students are aware of where the easy and difficult items lie in a test, the order in which the items are presented affects neither their attitude towards how they approached the test nor on their overall scores.

本文探討測驗中問題的次序是與學生得分有關係。一般人認為應把較易的項目安排在測驗開端，受試學生會因而產生正面的動機效應。但本研究結果卻是相反。學生問卷調查顯示學生雖留意到測驗裏難易項目的分佈情況，項目的次序卻沒有影響到他們做測驗的態度或他們的整體得分。

This paper investigates the extent to which the ordering of questions in an test has a bearing on students' scores on that test. Comments by authors of books on how to develop effective and well-constructed tests comment that one feature of a 'good' test is that items in a test ideally need to be arranged with easy items at the beginning of the test, leading to the more difficult items towards the end of the test. This, it is claimed is a motivating factor, a factor which encourages students to continue with a test which they might otherwise find difficult. Gronlund (1985, p. 137) proffers the following advice to test designers on item facility and how to arrange items in tests

> *"Except for a few items at the beginning of the test, for motivational purposes,* none of our items should be so easy that everyone answers it correctly".

It can be argued that it makes sense to order the items on a test from easy to difficult on a long placement test, on the basis that, on this type of test, students give up once they are past their own ability threshold level. This notion would, however, appear to have infiltrated to the extent that it has become an accepted principle of apparently good test design. Baker (1989, p. 51) in discussing item facility index, advises in a similar manner to Gronlund

> "This [item facility] index can be useful when deciding the order of items in a sub-test. *It is generally desirable to start the test with an easy item*".

Other authors of books who offer similar advice are, to name but two: Madsen, 1983, p. 182; Heaton, 1975, p. 177.

Previous research which has focused on these issues – that is, the extent to which the order of items in a test affects student performance, and increases student motivation — have produced mixed results. Certain researchers have claimed that item ordering has little effect on student performance (Klein & Bolus, 1983; Klimko, 1984). Other researchers have suggested that the picture is not so clear and that item ordering may indeed affect student performance (Hambleton & Traub, 1974; Tippets & Bensen, 1989).

One problem has been with the type of test administered. The types of multiple-choice tests which have been used have been reading tests or tests of content knowledge (Klein and Bolus, 1983; Balch, 1989; Cizek, 1991). Consequently, testees may have not necessarily completed the test in the order in which it was presented to them. They may well have answered questions in a more random order, on the basis of either questions which they found easiest or questions to which they felt they knew the answer. The current paper has attempted to avoid this design-fault by giving testees a listening test, which therefore must be answered in the order pre-set by the test designer.

Method

The study involves the Hong Kong Examination Authority's (HKEA) Certificate of Education (CE) English language examination (Syllabus B). Approximately 120,000 candidates take this examination in Secondary 5 (11th grade) when they are about 16-17 years old.

The test given to students consisted of three parts:

1. a 30-item multiple-choice listening test;
2. a 15-item multiple-choice cloze usage test;
3. a short questionnaire on the students' reactions to the test they had just taken.

Two different test batteries were assembled. To guard against the possibility of students having done either section of the test before, the 1985 multiple-choice cloze usage test was selected, as the multiple-choice papers from that year were not published. 30 listening items were then assembled from the 1991 and 1992 listening papers[1]. With an overall test mean of 50% as the target, items with facility ranges of between .24 and .81 with good discrimination values (generally > 0.3) were selected from the ten sessions of the 1991 and 1992 CE listening tests.

The cloze passage was included in both tests to provide an indicator of both groups' abilities vis-a-vis each other compared with the CE whole group figures, as well as to provide a reliability anchor against which the two groups' performances could be measured on the listening tests.

Two comparable groups in terms of ability level were then assembled from Form 5 classes in Hong Kong secondary schools which were just about to take the 1993 CE examination. The tests were run in five separate schools, with parallel classes in each school, one taking both the easy-to-difficult test (n=163) and another taking the difficult-to-easy test (n=179).

The previous CE whole group figures for the usage test and the listening test are presented in table 1 below.

TABLE 1
CE Whole Group Figures

	Usage test	Listening test
Mean	57%	51%
SD	16.2%	19.4%

The HKEA considers 50-55% as the optimal mean for the various papers of its English language examinations. Table 1 above shows that both the usage test and the listening test are close to these optimums, although the usage test had a slightly higher mean.

The results were then analysed, using the SPSS/PC+ statistical package, under the following parameters:

1. mean and standard deviations of both groups: to see how these compared vis-a-vis the CE whole group;
2. t-tests for the two groups;
3. chi-squares for the two groups' questionnaire results;
4. item analyses of the items on the two listening tests.

Results

The means and standard deviations of both groups are presented below in table 2.

TABLE 2
Test Means and Standard Deviations

	EASY-TO-DIFFICULT		DIFFICULT-TO-EASY		CE WHOLE GROUP	
	Usage	Listening	Usage	Listening	Usage	Listening
Mean	53%	51%	50%	50%	57%	51%
SD	20.5%	20.0%	19.6%	19.7%	16.2%	19.4%

As can be seen from table 2 above, the two groups' scores on both parts of the test are very close to the CE whole group figures: this demonstrates that the current sample is representative of the target population, from which the test papers were drawn.

T-tests were run betwen the two subsections of the test. No significance was revealed between either the students' scores on the usage test ($t(340) = -.59$, $p = .63$), as would have been expected, nor between the scores of the two groups on the DE and ED listening tests ($t(34)) = -1.34$, $p = .59$). There would therefore appear to be support for the first hyphothesis, that students' scores are not affected by the order of items on a test.

The questionnaire was then analyzed to see how students perceived the test in terms of how much they enjoyed taking tests, whether they felt the questions were easy or difficult, and what their expected scores were. If it is true that students are motivated by easy questions at the beginning of a test, or conversely demotivated by a lot of difficult questions at the beginning of a test, the two groups' questionnaire results would be significantly different. So that it was not obvious that the focus was on the listening items, questions on the questionnaire were posed about both the usage test and the listening test. Questions were generally laid out on a 4-point ordinal scale.

The results are presented in table 3 below.

TABLE 3
Questionnaire Results on Attitudes to Tests and Test-Taking

	EASY-TO-DIFFICULT					DIFFICULT-TO-EASY					Chi-square
	1	2	3	4		1	2	3	4		
	no			yes		no			yes		
Enjoy tests	17%	31%	43%	9%		21%	35%	36%	8%		2.07
Enjoyed U test	14%	35%	39%	12%		21%	29%	37%	13%		2.84
Enjoyed L test	21%	42%	26%	11%		29%	36%	27%	8%		3.19
	1	2	3	4		1	2	3	4		
	easy			diff		easy			diff		
U test easy/diff	4%	37%	47%	12%		7%	36%	47%	10%		1.92
L test easy/diff	4%	26%	49%	21%		2%	31%	50%	17%		2.36
	1	2	3	4	5	1	2	3	4	5	
	none	beg.	mid.	end	all	none	beg.	mid.	end	all	
Easiest U qs	19%	26%	25%	24%	6%	17%	19%	25%	33%	6%	4.65
Easiest L qs	27%	43%	18%	6%	6%	25%	19%	25%	24%	7%	37.27*
	<10	11-20	>20			<10	11-20	>20			
Expected U score	34%	62%	n/a			38%	62%	n/a			3.25
Expected L score	29%	60%	11%			26%	62%	12%			0.41

Notation
U = usage; L = listening; beg = beginning; mid = middle; diff = difficult; qs = questions
* $p < .0001$

As can be seen, response patterns were similar for both groups across almost all questions. Chi-squares values were non-significant except for the question where students were asked where the easiest questions were on the listening test $X^2 (4) = 37.27$, $p < .0001$).

The fact that chi-square has thrown up significance underlines the fact that the students are aware of which items on the test are difficult and which are easy. The same question asking where the easy usage questions were was not significant, revealing similar response patterns to the other paired sets of questions.

Item analyses were run for each of the two groups. These again revealed great similarity with the 1989 whole group, and with each other. The usage test item analyses, as would be expected given the similarity of the overall figures for both groups, were virtual mirror images of the CE whole group's: even down to the fine detail of which items had the highest or lowest facility values, and which discriminated better or worse — again a reaffirmation of the tremendous reliability of multiple-choice as a test type.

More diversity might have been expected with regard to the items on the listening test, given

that they were arranged as two different tests. The results, however, belie this, as can be seen from table 4 which presents the facility value for the 30 items.

TABLE 4
Item Facilities

Item No.	CE	DE	ED	Item No.	CE	DE	ED
1	.81	.73	.69	16	.41	.48	.34
2	.76	.82	.70	17	.39	.39	.33
3	.72	.72	.65	18	.38	.43	.37
4	.70	.61	.53	19	.39	.44	.41
5	.68	.58	.63	20	.37	.41	.33
6	.67	.63	.70	21	.34	.39	.43
7	.67	.63	.65	22	.33	.26	.31
8	.65	.48	.55	23	.33	.48	.49
9	.65	.61	.65	24	.31	.30	.31
10	.65	.61	.63	25	.31	.30	.39
11	.62	.61	.65	26	.29	.27	.33
12	.61	.70	.58	27	.28	.32	.31
13	.61	.70	.69	28	.27	.33	.33
14	.61	.68	.68	29	.25	.19	.21
15	.59	.57	.49	30	.24	.23	.32

Notation
CE = CE whole group
DE = difficult-to-easy group; ED = easy-to-difficult group

The items have been run on three different tests: the CE whole group in obviously random order, and the two tests in the current project. Given this, the correspondence among the three sets of items is very significant, with the two experimental groups' facilities very closely mirroring those of the CE whole group. A similar picture was obtained with the item point-biserial discrimination values. These are not presented here, but very comparable results were obtained across all three groups with high discrimination for each item. (One of the categories for the original inclusion of a particular item was the fact that they had good discrimination in the live examination.) The mean discrimination for the 30 CE listening test items was 0.37; for the easy-to-difficult and difficult-to-easy groups they were 0.42 and 0.41 respectively, demonstrating that the items are working well in discriminating among the better and less able students. As would be expected from the results presented so far, high correlations were obtained from the facility values; these are laid out in table 5 below.

TABLE 5
Correlation Coefficients Among the Sets of Items

	ED	CE
DE	.9278 p = .000	.9183 p = .000
ED		.9274 p = .000

Notation
CE = CE whole group
DE = difficult-to-easy group; ED = easy-to-difficult group

Correlations among all three groups are highly significant ($p < .001$), further underlining the fact that the orders in which the items have been presented to the students have not affected the results they have produced.

Discussion and Conclusions

The above data has presented three findings from which it must be concluded that the ordering

of items on a test affects neither students' performance on that test nor their motivation, or lack of it in how they approach the test:

1. Students' scores on the easy-to-difficult test were very comparable with those on the difficult-to-easy test, showing no favouritism to those who were presented with the easy items early on in the test.
2. Results of the questionnaire also showed no significance between either group in terms of their attitude towards test-taking or their expected scores, even though the data did reveal that students were able to pinpoint where the easy or difficult questions lay in their respective tests.
3. Item analyses for the two tests revealed very similar results for the same item. It can be assumed that an item's facilitiy and discrimination values will generally not be affected by where that item may be located within a test.

One criticism of the current methodology might, however, be that since the current tests were administered under classroom conditions, it might be difficult to generalise the findings to public examinations, where the stresses of a real-time examination are more apparent. The current study has attempted as far as possible to take this factor into account by having students take the current tests just before their actual 1993 public examinations. They had been doing practice tests and were therefore in a state of 'exam-readiness'.

The current study investigated the hypothesis that the order of questions within a test does not affect students' performance. A listening test was selected as the operant test type on the basis that the order in which the items are presented cannot be circumvented since testees have no alternative but to switch their attention from one item to the next as they hear the items on the tape. The opinion of the current author is that the fact that the test type was a listening test is immaterial; the results should hold true for any test type.

Another related matter concerns the extent to which it many be argued that the current research is only applicable to multiple-choice testing. As a percursor to the current study, the author conducted a pilot study which looked at a 12-item matching exercise from another Hong Kong English language examination — a reading test. As it was a pilot study, the sample was smaller (n=30 in each sample group). Similar results were obtaianed to those described in the current study. It was found that item position in the test on the basis of item facility affected neither the respective test mean nor individual item facility, which were again very similar to the Hong Kong live examination figures. From this it can be assumed that the results from the current study also hold for non-multiple-choice tests.

The final word to test designers and teachers is therefore: it does not matter the order in which you present test items to your students or candidates. Performance and motivation are unaffected by test item order.

Note

[1]The CE listening test consists of six sections, with a mixture of short items and extended listening pieces. The item type that was selected for examination were the 'skim and scan' type of item. Here students see a number of similar pictures, charts or diagrams. After listening to a short conversation or description, they have to decide which picture is being discussed.

The CE listening test is run in 5 parallel sessions. Only two of these are published each year.

References

Balch, W. R. (1989). Item order affects performance on multiple-choice exams. *Teaching of Psychology, 16*(2), 75-77.

Baker, D. (1989). *Language Testing*. London: Edward Arnold.

Cizek, G. J. (1991, April 4-6). *The effect of altering the positions of options in a multiple-choice examination.* Paper presented at the Annual Meeting of the National Council on Measurement in Education. (Chicago, April 4-6, 1991).

Dambrot, F. (1980). Test item order and academic ability, or should you shuffle the test item deck? *Teaching of Psychology, 7*(2), 94-96.

Gronlund, N. E. (1985). *Measurement and evaluation in teaching, 5th edition.* New York: McMillan.

Hambleton, R. K., & Traub, R. E. (1974). The effects of item order on test peformance and stress. *Journal of Experimental Education, 43*(1), 40-46.

Heaton, B. (1975). *Writing English language tests.* London: Longman.

Klein, S. P. (1983, April 12-14). *The effects of item sequence on bar examination scores.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

Klimko, I. P. (1984). Item arrangment, cognitive entry characteristics, sex, and test anxiety as predictors of achievement examination performance. *Journal of Experimental Education, 52*(4), 214-219.

Madsen, H. S. (1983). *Techniques in testing.* Oxford: Oxford University Press.

Tippets, E., & Bensen, J. (1989). The effects of item arrangement on test anxiety. *Applied Measurement in Education, 2*(4), 289-296.

Author

David CONIAM, Lecturer, Faculty of Education, Chinese University of Hong Kong, Shatin, New Territories, Hong Kong.