JOHNSTON ANDERSON

SEX-RELATED DIFFERENCES ON OBJECTIVE TESTS AMONG UNDERGRADUATES

ABSTRACT. Objective tests, such as multiple-choice and relationship analysis questions, are often used in mathematics to assess certain ranges of skills and abilities. This article examines the performance of men and women students in mathematics at university level and shows that there are significant differences between the sexes.

1. INTRODUCTION

Much research has been carried out in the last 20 years or so into influences which might cause or explain sex-related differences in mathematics performance, at a number of different ages and levels, notably by Fennema (1977, 1978, 1979, 1985), Husén (1967), Leder (1980, 1985), Maccoby and Jacklin (1975), Sherman, J. A. (1977) and Sherman, S. W. (1974). These, and other recent sources such as APU (1985), Burton (1986) and DES (1985) show that there are considerable differences in the proportion of boys and girls in the top ability bands, that (in the UK) more boys get better grades at A-level and that boys and girls respond differentially to multiple-choice questions. However, most of the studies cited above tend to refer to high-school pupils up to the age of 18. The present article, by contrast, focusses exclusively on the performance of *university* students majoring in mathematics, to see if these might display similar differences. One might suspect not, given the selective nature of university entrance in the UK and the consequent high achievement levels required to gain entry; moreover, one might further presume that both men and women entrants are equally motivated towards the course of study they have chosen. It was therefore a surprise to find that, in a recent example of an objective test, the mean score of the men was significantly higher than that of the women. This prompted a re-examination of the results of other objective test questions given previously, including those discussed in an earlier article (Anderson, 1979).

In all, five examples of test question are considered, which were attempted by four different groups of students. These four groups of students comprised the entire first-year entry (from four different academic years) to specialised mathematics degree courses at this institution. The objective test questions were part of their end-of-first-year examination. As mentioned

Educational Studies in Mathematics **20**: 165–177, 1989.

earlier, students entering the course require high qualifications at advanced (A-) level examinations. While grade A in mathematics is attained by about 12% of the country-wide entry and grade B by the next 15% or so (grades A to E being pass grades, achieved by about 70% of the entry), among the four groups of students considered, all but about 3% of each sex achieved at least grade B in mathematics. In the case of the last example, discussed in Section 5, no fewer than 54 of the 56 men and 22 of the 24 women gained grade A, the remaining four gaining grade B. Among the other three groups (113, 118 and 101 students respectively), the proportions of men and women gaining grade A was less (typically between 60% and 75%) with slightly more men than women gaining grade A, though the difference was not significant. The assessment of quality is further complicated by the fact that some students attempt two mathematical subjects at advanced level, usually taking "Further Mathematics" in addition to "Mathematics". Almost half of all the entrants over the four years achieved the two highest double grades (AA or AB). In two of the four years, more women than men did so, in the other two years less, though in no case was the difference significant. In summary, the four populations considered form a narrow and homogeneous stratum at the top end of the school ability range.

Performance on "standard" first-year university examination papers also revealed no significant difference between the sexes, but, on the objective tests considered in this article, there do, by contrast, seem to be noticeable and significant differences suggesting that women are, on the whole, less successful than men on tests of this kind on this subject material. This has implications for the wider use of objective testing, for example, in computer-marked tests (which may well become more common). This is a point also made, though with particular reference to physics and chemistry, by Harding in Kelly (1981, pp. 198, 284).

Different formats are possible and, in the next four sections, examples of the most frequently used are examined. The tests, or questions, were used to assess factual knowledge, relationships between concepts introduced during a course of instruction and the low-level understanding of concepts and their consequences; they were not (and are not normally) employed to assess higher-order skills such as problem solving or proof.

2. MULTIPLE-CHOICE QUESTIONS

The first example is of a standard multiple-choice question, containing six items. Each item describes a mathematical situation, followed by five conclusions, of which the student is required to choose one (and only one). Full details with some general comments on the students' performance can be found in Anderson (1979), Section 4, though this deals with the population as a whole and does not discriminate between the sexes. A correct answer scores +4 and a wrong answer −1, with no penalty for leaving all options blank. The question was attempted by the complete first-year entry of 55 men and 58 women, whose overall scores on the whole paper averaged 51.10% and 51.23% respectively, with a similar spread and range of marks.

However, on the multiple-choice question (marked out of 10), the men scored an average of 4.01 and the women 3.31. Although the absolute difference in marks is small (seven-tenths of a mark out of a paper total of 100), the *relative* difference is quite high; the women students' average mark on this question is only 83% of the men's average mark. Over a more substantial (or even complete) paper in this format, this could produce significant differences disadvantageous to women students and, in examinations where grading is important (such as English 'A' levels), quite substantially affect the grades awarded.

A worse performance can arise either because the student gets parts wrong or because the student leaves parts unanswered ('blank'). In the case of this multiple-choice question, few candidates left parts unanswered and the women students did worse because they got more answers incorrect (Table I). This is not always the case as we shall see.

However, closer examination of the items in this question shows that, while a higher percentage of men *on every item* were successful, in only one item (namely (c)) was this difference very significant. Even on those items ((d) and (e)) most badly done, there was little difference between the sexes, the men being marginally the more successful. Table II shows the percentages of men and women getting the different items correct and incorrect, respectively (parts left unanswered not included). There was no apparent relationship between the sex of the student and which particular wrong option was chosen.

TABLE I

	Correct (%)	Incorrect (%)	Blank (%)
Men	51.2	44.6	4.2
Women	45.4	49.4	5.2

TABLE II

	Items correct (%)					
	a	b	c	d	e	f
Men	69	71	67	35	18	49
Women	67	64	50	31	14	47
	Items incorrect (%)					
	a	b	c	d	e	f
Men	31	27	33	53	73	49
Women	28	36	47	60	74	52

3. RELATIONSHIP ANALYSIS QUESTIONS

The example used (see Anderson, 1979, Section 5) also contributed 10% to the same examination paper. On this occasion, students are presented with two statements, p and q, and asked to identify which, if any, of the following four relationships apply. A: p implies q (but not vice-versa), B: q implies p, C: p and q are equivalent, D: p implies 'not q'. A fifth option is E: none of the above relationships apply.

Ten items were used and the scoring is as in the previous example. Table III shows the results. As in the previous case, although the women are, on average, only three quarters of a mark worse, their average score is about 86% of the corresponding men's score.

However, unlike the example above, a more detailed examination of the ten items in this question does not show a consistent pattern. Table IV shows the percentages of men and women, respectively, giving (a) correct, (b) incorrect, and (c) blank responses.

We see that on items (b), (i) and (j), the men collectively got substantially more correct, and fewer incorrect, than the women; on items (g) and

TABLE III

Responses to relationship analysis questions

	% Correct	% Incorrect	% Blank	Net score (out of 10)
Men	58.6	28.7	12.7	5.14
Women	52.1	31.9	16.0	4.41

TABLE IV

	a	b	c	d	e	f	g	h	i	j
Correct responses (%)										
Men	80	78	47	27	56	76	60	69	36	55
Women	79	57	52	26	53	84	53	66	17	33
Incorrect responses (%)										
Men	16	18	40	44	42	13	36	27	7	44
Women	17	36	34	48	43	9	33	24	17	57
Blank responses (%)										
Men	4	4	13	29	2	11	4	4	56	1
Women	4	7	14	28	4	7	14	10	66	10

TABLE V

Number of 'blank' answers:	0	1	2	3	4	5	6	7
Men	21	14	8	8	4	0	0	0
Women	15	21	11	4	2	2	1	2

(h), by contrast, not only did a higher percentage of men get the correct answer but a higher percentage also got an incorrect answer; in these cases, as in (i) and (j), a considerably greater percentage of women left this item unanswered. It is difficult, without individual interview, to discern reasons why performance might be different in these cases, although in the case of item (i), the complicated appearance of "q" was sufficient to deter a large proportion of *all* students from attempting the item, and may support the view expressed by Fennema and Peterson (1985) that males do better, and females worse, on tasks of high cognitive complexity.

Of some interest also is the distribution of blank responses by sex. Table V shows the number of students leaving different numbers of items unanswered. Thus, while 38% of men attempted all ten items, only 26% of women did. Furthermore, the average number of blank responses was 1.27 for men but 1.60 for women, and it was women students who produced the most extreme behaviour (five attempting 5 or fewer items).

There may be some support in this for Sherman (1974), who points out that greater self-confidence leads to more items being attempted and, when an 'I don't know' option is available, that it is those who are correct less often who say 'I don't know' more often. Sherman concludes that sex differences in correct response percentages for many of the exercises she conducted at adult level can be explained almost completely by differences in the usage of 'I don't know' options; opting for 'I don't know' may be the result of a fear of risk-taking or a fear of being wrong or even a lack of motivation. Although no such option was available in the question under discussion, the implication in the rubric that no penalty would be incurred for leaving an item unattempted may have had a similar effect. The issue of self-confidence is one we shall return to in Section 6.

4. TRUE/FALSE QUESTIONS (I)

The type of questions used here involved a number of blocks, each of five items, each block being devoted to a single theme. In contrast to the previous examples, instead of selecting one (and only one) statement from five available, the student is required to determine, for each item, whether it is true or not. This is a harder task than a standard multiple choice question, since a correct identification of the true answer in a multiple-choice item implies that all options are untrue, whereas in the true/false format, each statement stands alone and must be treated independently of the others. The scoring method for such questions gives a correct answer +1 and an incorrect answer −1, with no penalty for leaving an item blank.

In the first example below, there were nine blocks (i.e. 45 items in all). Sixty men and 58 women participated in the test (the total first-year entry but from a different year from that considered above). The outcome is shown in Table VI.

The average 'score' (out of 10) for men was 3.91 and for women was 3.33, so that the women's average mark was about 85% of the men's average mark. The proportion of incorrect answers was almost identical overall, the difference in scores being attributable to the lower number of blank responses by the men. The distribution of the number of blank answers between men and women was significantly different. In the case of men, while the number of blank answers ranged from 0 to 26, the mean and median were, respectively, 6.68 and 5.5 and, in addition, 15% of men

TABLE VI

True/false test

	Correct (%)	Incorrect (%)	Blank (%)
Men	62.1	23.0	14.8
Women	55.8	22.5	21.7

TABLE VII

	Number of blank responses								
	0–3	4–7	8–11	12–15	16–19	20–23	24–27	28	Σ
Men	22	15	13	5	3	1	1	0	60
Women	9	14	17	6	8	1	2	1	58

attempted *all* the items. For the women students, on the other hand, the number of blank answers ranged from 0 to 28, but the mean and median were 9.76 and 9 respectively. Moreover, for women, the most popular number of blank responses was 9, attained by 12% of the population, while only 6% answered all the parts. The difference is illustrated in Table VII where the blank scores have been grouped for ease of presentation.

There are, however, some interesting variations when we examine individual items. One such item involved five numbered assertions about a series of positive terms (assertions such as "(1) if Σa_n converges, so does Σa_n^2; (2) Σa_n^2 diverges only if Σa_n^4 diverges").
Students then had to mark each of the five statements

(i) (1), (4) and (5) are true
(ii) (2) is false
(iii) (1), (3) and (4) are true
(iv) (1) and (5) are true
(v) none of the above four statements is correct

true or false. This item, not surprisingly, wasn't very popular. Nonetheless, while it produced a large percentage of blank responses per item it revealed two significant facts. On the other hand, it was the women who were notably the more cautious on all five parts (the complexity factor again?); however, the greater impetuosity of the men did not turn out to their disadvantage (Table VIII).

If we compare the difference between the percentages of correct and incorrect answers ("C-I" in Table VIII), we see that, with the exception of part (ii), where twice as many men as women gave incorrect answers, the men's net results were better than the women's by between 22 and 37 percentage points ("Excess M-W" in the table). Part (i) was also somewhat of a surprise in that it produced the only negative net balance (−17%) of correct against incorrect answers. There seems no obvious reason why women got this wrong so often – in cases of serious doubt, most students (including women) tend to leave options unanswered.

TABLE VIII

		Correct (%)	Incorrect (%)	Blank (%)	C-I	Excess M-W
(i)	M	50	30	20	+20	+37
	W	24	41	35	−17	
(ii)	M	37	25	38	+12	−4
	W	28	12	60	+16	
(iii)	M	58	15	27	+43	+22
	W	40	19	41	+21	
(iv)	M	55	23	22	+32	+22
	W	29	19	52	+10	
(v)	M	70	8	22	+62	+30
	W	39	7	54	+32	

Several other items, or parts of items, on this question produced high percentages of blank responses, but the differences in performance of men and women was more varied. Another item concerned complex numbers: "Given that the complex number z lies in the second quadrant of the Argand diagram, it follows that

(i) z^2 lies in the fourth quadrant
(ii) $\arg(z - z^*) = \frac{1}{2}\pi$
(iii) z^* lies in the third quadrant
(iv) $\arg(z + z^*) = 0$
(v) $|e^z| \leqslant 1$."

As will be seen from Table IX, women students once again tended to be more cautious, markedly more so than the men in parts (i) and (v). The willingness of the men to commit themselves, although paying dividends in part (i), turned out to be misplaced in part (v). The proportion of women getting part (iv) incorrect was unexpected.

Finally, in an item which concerned formal logical relationships ("given that P is true if Q is true and that Q is false only if R is false, it follows that..."), there were again parts with a high abstention rate. In two such parts (which we shall call (i) and (ii)), the conclusions to be decided were

"(i) a necessary condition that R be true is that P be true"

and

"(ii) R is true except when P is false."

The outcomes are shown in Table X.

TABLE IX

		Correct (%)	Incorrect (%)	Blank (%)	C-I	Excess M-W
(i)	M	80	12	8	+68	+35
	W	52	19	29	+33	
(ii)	M	43	17	40	+26	+2
	W	43	19	38	+24	
(iii)	M	63	15	22	+48	+2
	W	60	14	26	+46	
(iv)	M	33	32	35	+1	+26
	W	15	40	45	−25	
(v)	M	21	32	47	−11	−17
	W	16	10	74	+6	

TABLE X

		Correct (%)	Incorrect (%)	Blank (%)	C-I	Excess M-W
(i)	M	28	50	22	−22	−19
	W	33	36	31	− 3	
(ii)	M	34	33	33	+ 1	−18
	W	38	19	43	+19	

These present a somewhat different picture from most of the earlier evidence. The women abstained more often than the men (as has been consistently the case in previous examples), but in these two cases, both more women got the answer correct and many more men than women got the answer wrong, resulting in large negative values for M-W in the table. This supports some other evidence found by the author suggesting that, on questions concerning 'pure' logic, women often perform better than men. It may be the case that this kind of question relies more than others on verbal reasoning and linguistic facility, areas in which females have been shown, by a number of sources in the references, to be better than males.

In a second example of this type of question, with slightly easier items, but with a different population (of 47 men and 54 women, again a complete first-year entry with similar qualifications), the mean score of the men was 4.98 (out of 10) and that of the women 3.98 (in this case about 80% of the men's mean score). The men averaged 69% correct, 19% wrong and 11%

blank, against 62%, 22% and 16% respectively for the women. Patterns of abstention between the sexes were similar to that of the previous true/false test.

5. TRUE/FALSE QUESTIONS (II)

The way an instruction is expressed may have a considerable effect on the number of blank responses. In the next example, no mention was made of penalties for incorrect answers. The question, a shortened version in the style of the earlier questions above, contained only nine items and was more directly related to the content of the course:

"Mark each of the following true or false. For those that you believe to be false, give a counterexample; you are NOT required to prove those you believe to be true.

(a) $(ax+b)^2+4ac-b^2>0$ for all $x \in \mathbb{R}$ only if $b^2<4ac$.
(b) If the fixed number y satisfies the inequality $y<7+x$ for every $x>0$, then $y \geqslant 7$.
(c) If, for every $x \in A$, $x \leqslant \alpha$, then $\alpha = \sup A$.
(d) If $(x^2+2x+3)+y>0$ for all $x \in \mathbb{R}$, then $y>0$.
(e) If the real sequence (s_n^2) is decreasing, then it converges to zero.
(f) If $\sum_{n=1}^{\infty} a_n$ converges, then $a_n \to 0$.
(g) If $f'(a)$ exists, then f is continuous at a.
(h) If $f'(a)=f''(a)=0$, then a is a point of inflection of f.
(i) If z and w are complex with $|z|>1$ and $|w|>1$, then $|z-w| \leqslant |z^2-w^2|$."

TABLE XI

	Correct responses (%)								
	a	b	c	d	e	f	g	h	i
Men	79	77	82	70	63	100	98	68	30
Women	83	71	50	79	46	83	92	63	0
	Incorrect responses (%)								
	a	b	c	d	e	f	g	h	i
Men	21	23	18	29	36	0	2	29	68
Women	13	29	50	17	54	17	8	33	92

Significantly, the number of abstentions on this question was very small indeed (1% for men, 2.3% for women). The population attempting the question consisted of 56 men and 24 women, almost all of whom had very high qualifications in mathematics. Yet again, however, the women performed worse overall, averaging only 63% correct against 74% by the men.

There being no penalty for incorrect answers, the women's mean score of 6.3 is 85% of the mean score (7.4) of the men. Examination of the nine items individually reveals some strange curiosities (Table XI), the most extraordinary being the failure of any women at all to get part (i) correct. However, the divergence in scores between men and women in parts (c) and (e) is also not easily explained.

6. CONCLUSION

The examples above, involving four different groups of students and four different types of objective test, demonstrate an unexpected consistency in the relative performances of men and women undergraduates on these questions. It is a surprise, and no doubt fortuitous, that the mean score of women on all five tests was between 80% and 86% of the corresponding men's score. But while the uniformity of these figures may be accidental, the fact that women were consistently worse seems not to be. This can be partly explained by a higher level of abstentionism, suggesting that women are less likely to gamble when they are really unsure of the answer. Annice *et al.* (1988) report the results of a survey of a very large (1.15 million) population of Australian school pupils which showed that, in multiple-choice tests, girls were less willing to guess answers. (See also *New Scientist* (1988).) Indeed, Sherman (1974) claimed that much of the sex-difference in science performance as reported by the National Assessment of Educational Progress is an artefact of sex-differences in test-taking behaviour rather than intrinsic ability, a theme also explored by Arzi (1985) and Benedictis *et al.* (1982). Greater self-confidence in their mathematical ability is noted among boys in grades 6–12 by Husén (1967), by Fennema and Sherman (1977, 1978), and by Adams (1985). These differing degrees of self-confidence ("the most persistent and pervasive finding") may indeed be a significant factor among such mathematical high-fliers as those considered here. Of course, the guesses students make are not random but rather 'educated' guesses, illustrated by the fact that some distractors are much more popular than others, but we note that women also performed worse than men in those cases where abstentionism is low (Sections 2 and 5). This may offer support to the view expressed by Leder (1980) that "differences increase as the level of examinations taken increases and is particularly marked when above average performance is considered". This raises the question whether, among the students considered here, sex-related differences were more pronounced among 'above-average' men and women. Because of the homogeneity of their entry qualifications, it is not easy to define 'average' and 'above-average' satisfactorily; one possibility is to use eventual degree-class as a measure, but it must be remembered that degree class is heavily determined by conventional examinations and very little by objective tests. Nevertheless, if one looks at the results of the questions above and compares these for men and women who achieved 'good' degrees and also for those gaining average or below-average degrees, then in both cases the comparisons were not significantly different from the results of the sex-groups taken as a whole, described in the tables above: women still scored, on average, between 80% and 90% of the corresponding men's mark. A further point to note is that mathematics degree courses are "co-educational" and authors such as Shuard (1982) and Harding (in Kelly, 1981) argue that females seem disadvantaged in a mixed (school) setting.

The results above are particularly striking because they apply to different populations attempting different questions at different times. Provided that the proportion of objective testing within the examination structure as a whole does not become too large, there may be little harm done (and objective tests can have considerable value as a diagnostic tool), but it should be recognised that a problem may exist. There seems to be enough evidence to suggest that the extensive use of such tests may be detrimental to women and should be used with caution.

REFERENCES

Adams, R. J.: 1985, 'Sex and background factors: Effect on ASAT scores', *Australian Journal of Education* **29**, 221–230.

Anderson, J. A.: 1979, 'Objective testing in elementary analysis', *Educational Studies in Mathematics* **10**, 227–243.

Annice, C. *et al.*: 1988, 'Gender differences in the Australian mathematics competition', Paper presented to Annual Meeting of *Australian and New Zealand Association for the Advancement of Science* (ANZAAS).

APU: 1985, 'A review of monitoring in mathematics 1978 to 1982', *Assessment of Performance Unit*, HMSO, London.

Arzi, H.: 1985, 'Gender differences in science achievement; a matter of knowledge or confidence?', Paper presented to Annual Meeting of *American Educational Research Association.*

Benedictis, T. *et al.*: 1982, 'Sex-related differences in science: I don't know', Paper presented to Annual Meeting of *American Educational Research Association.*

Burton, L.: (ed.): 1986, *Girls into Maths Can Go*, Holt Education, Holt, Rinehart and Winston.

DES: 1985, *Statistics of Education, School Leavers CSE and GCE, England 1983*, Department of Education and Science, HMSO, London.

Fennema, E.: 1979, 'Women and girls in mathematics – equity in mathematics education', *Educational Studies in Mathematics* **10**, 389–401.

Fennema, E. and P. L. Peterson: 1985, 'Autonomous learning behaviour', *Educational Studies in Mathematics* **16**, 309–311.

Fennema, E. and J. A. Sherman: 1977, 'Sex-related differences in mathematics achievement and related factors', *American Educational Research Journal* **14**, 51–71.

Fennema, E. and J. A. Sherman: 1978, 'Sex-related differences in mathematics achievement and related factors: a further study', *Journal for Research in Mathematics Education* **9**, 189–203.

Husén, T.: 1967, *International Study of Achievement in Mathematics*, Almqvist and Wiksell, Stockholm.

Kelly, A. (ed.): 1981, *The Missing Half*, Manchester University Press.

Leder, G.: 1980, 'Bright girls, mathematics and fear of success', *Educational Studies in Mathematics* **11**, 411–423.

Leder, G.: 1985, 'Sex-related differences in mathematics: An overview', *Educational Studies in Mathematics* **16**, 304–309.

New Scientist: 1988, 'Girls less willing to guess than boys', **118**, No. 1614, p. 34.

Sherman, J. A. and E. Fennema: 1977, 'The study of mathematics by high school girls and boys: Related variables', *American Education Research Journal* **14**, 159–168.

Sherman, S. W.: 1974, Multiple choice test bias uncovered by use of an 'I don't know' alternative, ERIC document ED 121824.

Shuard, H. B.: 1982, *Mathematics Counts*, W. Cockcroft, ed. HMSO, London.

Department of Mathematics
University of Nottingham
Nottingham NG7 2RD
UK