

Gender Bias and Test Norms in Educational Selection

Harvey Goldstein
University of London
Institute of Education

The Equal Opportunities Commission considers it discriminatory deliberately to balance the gender composition of schools or streams: "any allocation made should be solely on the grounds of ability". In particular, where tests are used, standardised scores should not be based on separate sex norms.

Some LEA's appear to have concurred readily with this view, while others have been more reluctant. A particular instance of the latter is where the existence of the same number of places for boys and girls in grammar schools has led LEA's to use separate sex norms for tests, leading to identical distributions of standardised scores for each sex, but also, because girls tend to have higher 11+ raw test scores, resulting in some girls being excluded from grammar schools who had higher scores than some boys who had been selected. The paper explores the arguments which have been used to justify this procedure, and the implications for assessment in general, including public examinations.

Background

According to Wilby (1983), a then current court case threatened the future of single-sex grammar schools because an LEA was operating separate sex norms for its 11+ selection tests. What was happening was that, because girls obtained higher average scores than boys but then had their scores adjusted downwards to equalise the sex distributions, some girls were being refused admission to grammar school despite having higher raw scores than some boys who had been selected. Wilby points out that: "The case also has implications for councils with mixed grammar schools . . . a single pass mark . . . means that girls get a disproportionate number of places and many teachers feel the system is unfair to boys because their early development lags behind girls."

The case was being brought by the parents of the girl, with the backing of the Equal Opportunities Commission (EOC). The commission, in its guidelines (1982) states that "any allocation made (to schools or streams) should be solely on the grounds of *ability*" (my italics) and that separate sex norms should not be used. On

the face of things, if the Commission's view is upheld in law, it would seem clear that this LEA and others with similar practices, would be obliged to alter them, resulting in a higher proportion of girls entering grammar schools. In the remainder of this paper I will examine the arguments which surround this issue and relate them to some issues basic to testing and selection.

Sex Differences

Several studies using tests similar to those employed in 11+ selection (e.g. Douglas et al. (1968), Yates & Pidgeon (1957), Fogelman et al. (1978)) have found similar patterns of sex differences at the age of 11 years. Broadly speaking, for achievement in Mathematics and Reading the average differences are small, while for both verbal and non-verbal reasoning tests the girls have higher average scores than boys. By the age of 16 years the boys show distinctly higher mean scores for Mathematics and Reading but the girls retain their higher mean score on the reasoning tests.

Such findings, of course, are specific to the tests used and an immediate issue is whether, in some way, existing tests are 'biased'. Pidgeon (personal communication) reports an item analysis for two reading tests, one of which slightly favoured boys while the other slightly favoured girls. He concluded that these differences were due to a different balance of items favouring one or other sex and that appropriate selection of items could have yielded a test with any desired bias towards either sex; moreover that relative changes with age could also be manipulated.

It would be interesting to speculate on how existing tests have come to exhibit these differences, and whether test constructors have, in some way, built in particular sex biases as a result of their procedures. There is certainly evidence of how current cultural assumptions and expectations did influence test construction historically (Gould, 1981). Given the manner in which many new tests are validated by requiring them to correlate highly with existing tests, the persistence of historical biases is unsurprising. Nevertheless, the existence of the possibility of designing tests tailored to favour one sex, raises a number of problems directly relevant to the EOC judgement. Indeed, we may apply the same arguments to any identifiable groups, such as black and white children or those from different social groups, and there have been

attempts in the USA, for example, to devise tests which favour Blacks over Whites. It is worth noticing that this kind of manipulation of test items is quite different from procedures, as used at Educational Testing Services and elsewhere, aimed at producing sex and race-fair tests by eliminating negative stereotypes, derogatory references, etc. Even following such procedures, tests are still open to manipulation. I shall return to this issue after considering the principal arguments advanced in support of separate sex norms on the basis of existing testing using verbal or non-verbal reasoning tests.

Separate Sex Norms

Advocates of separate sex norms in selection argue that the relative superiority of girls over boys at the age of 11 diminishes or is reversed by 16, and is due to the earlier 'maturity' of girls. Thus, since this advantage is a temporary phenomenon, it should be allowed for and this is most conveniently done by selecting equal numbers of girls and boys. There are, however, a number of difficulties with this argument.

First, while the sex difference does change over time to favour boys in achievement tests, it appears not to do so for verbal and non-verbal reasoning tests. Since the latter are often the principal tests used in selection the maturity argument would seem to have little support. Secondly, there is no real educational justification given for selecting equal numbers of boys and girls. If we are to base an 'adjustment' on subsequent performance then, depending on the assessment used, this would not lead in general to equal proportions. Thirdly, subsequent performance is an unreliable guide, influenced as it is by the selection process itself, or by such factors as sex stereotyping and differential subject choice. Adams (1984), on the basis of a large scale study in Australia, concluded that "direct sex differences were found to be slight. This indicated that biological differences might be of relatively little importance in comparison with the more powerful societal and attitudinal factors". Fourthly, the idea of 'maturity' is borrowed from physical maturity, where in terms of bone maturity and size, girls are ahead of boys at the age of 11. As Tanner (1963) points out, however, measures of intellectual maturity are not available so that the analogy with physical development is improper. The only practical operationalisation of the term 'maturity' is with reference to changing performance differences over time, and there is no reason why these should be ascribed to some 'innate' difference between the sexes rather than, say, to factors operating within schools or society. In fact, the typical justification for the use of verbal and non-verbal selection tests is redolent of the old and

largely discredited view that IQ and reasoning tests in general reflect innate and unchanging abilities; a view which accords with the lack of change in the sex difference during the secondary school period, but which then hardly allows such tests to be used as measures of relative maturity. Finally, if one does accept the maturity argument, it is pertinent to ask why one should be content with adjusting only for sex differences. Thus, for example, as the number of other children in a child's household increases there is a relative decrease in reading and mathematics test scores between 11 and 16 years (Fogelman et al., 1978). It would seem quite plausible to argue that children from large households mature earlier and that this should be compensated for by using separate 'household size' norms. No-one seems seriously to have suggested this for 11+ selection although in the context of compensating for disadvantage, the issue may be very relevant.

In short, the major arguments in support of separate sex norms for 11+ selection have no sound rational justification and continued existence of the practice therefore would have to be justified on political and social grounds.

Conclusions

The wording of the EOC guidelines quoted in the introduction assumes that 'ability' (or attainment, or achievement) is theoretically recognisable and empirically measurable. Both these assumptions, however, are tenuous, but they raise the interesting question of whether it is legitimate to formulate a theoretical description of achievement which explicitly includes a sex difference or a difference between recognisable groups in general. Such a difference might involve the whole distribution of achievement scores or grades rather than simply a mean difference.

We can see that there may be a legitimacy if we try to imagine a formulation which requires no differences at all for this would then be making the assumption of equivalence between all groups. It seems therefore that some notion of expected or desired group differences (even if these are zero) is a necessary component of any theoretical model. Given such a choice, the problem becomes one of deciding which value system should determine the choice and how to deal with arguments such as those concerning compensation for disadvantage in admission to schools, higher education, employment and so forth.

In the UK this debate has not been as obvious as in the USA, but the issues remain. For example, if an LEA wished to comply strictly with EOC guidelines it should have little difficulty in finding somebody to produce a test for it which resulted in equal mean scores for boys and girls, so

having the same practical result as using separate sex norms but without falling foul of the law. Likewise, there seems to have been a movement in recent years by the public exam boards towards multiple choice questions. Yet there is evidence that such questions tend to favour boys as against girls (Murphy, 1982), so that this policy by the boards might be said to discriminate against girls, although the boards' motivations are, for example, financial rather than sexist.

If we accept that equality of outcome is both legitimate and desirable and so wished to have an exam which produced equal score or grade distributions for boys and girls, then we might well be able to achieve this by careful choice of question format, content etc. There are, of course, considerable difficulties in the way of achieving such an end, one being that entry rates for different exams differ markedly. Nevertheless, the resulting examinations might do much to encourage the sexes to participate more equally in certain subjects and the resulting effects on teaching and curriculum would be a rather interesting example of an assessment led pedagogy. As suggested earlier, gender is not the only characteristic one might wish to treat in this way, although it may well be easier and more socially acceptable to promote equality of assessment for the sexes than, say, for different social or ethnic groups. Yet, even after carrying through a procedure to eliminate, say, ethnic differences in order to equalise achievement, and even if this were technically feasible, it is somewhat difficult to imagine a consensus of agreement on the desirability, of course, attempts of this kind are anyway predicated upon the 'equality of outcome' assumption and there is a great deal of ideological discussion to be had on that issue. Needless to say, these arguments are equally relevant to current developments in assessment such as graded testing, and given the early stage of development of these initiatives it might be appropriate for those concerned to give some thought to the problems.

If we relinquish the notion of verbal and non verbal tests as measuring some kind of 'permanent' or 'innate' quality which achievement tests do not, there is no good reason why the latter alone should not be used for 11+ selection. In this case the girls would not in general surpass the boys. In fact, it would not be too difficult for an LEA which wanted to do so, to provide a coherent 'educational' rationale for using as selection instruments a collection of existing tests to produce any desired balance of the sexes. It would be an interesting point of law as to whether such a procedure would be considered inadmissible under the Act. In

reality, of course, there are no choices which are purely 'educational' set outside of a socio-cultural context. Cultural assumptions and expectations, organisational constraints and explicit ideologies all play a part in the measurement of achievement, and in inferring individual potential. Whenever selection occurs, such factors will influence the relative 'success' of certain groups over others, and there is still a great deal to be learnt about this.

As far as the 11+ is concerned, one resolution is to abolish selection at that age, and insofar as this would resolve some of the difficulties, it constitutes another argument in favour of comprehensive secondary schooling. The general issue, however, will not go away as easily so long as selection remains important elsewhere in the educational system, and the EOC might usefully turn its attention, for example, to the examination bodies. Thus, the apparent fairness and simplicity of the EOC guidelines advising against the use of separate sex norms for selection, hides a much more complicated and difficult problem. Mere paper compliance with the guidelines is no guarantee of effective change of practice at the age of 11 or at other stages of the educational system.

Acknowledgements

I am most grateful to the following for helpful comments: Karen Clarke, Caroline Gipps, Jim Houston, Douglas Pidgeon, Desmond Nuttall, Ray Sumner and Gaby Weiner. An earlier version of this paper was read at the BERA Annual Conference, Sheffield, 1985.

References

1. Adams, R.J. (1984) **Sex Bias in ASAT?** ACER Research monograph no. 24, Australian Council for Educational Research, Hawthorn, Victoria.
2. Douglas, J., Ross, J.M. and Simpson, H.R. (1968) **All Our Future**. London, Peter Davies.
3. Equal Opportunities Commission (1982) **Do You Provide Equal Opportunities?** Manchester, EOC.
4. Fogelman, K., Goldstein, H., Essen, J. and Ghodsian, M. (1978) Patterns of Attainment, **Educational Studies**, 4, 121-130.
5. Gould, S.J. (1981) **The Mismeasure of Man**, New York, W.W. Norton.
6. Tanner, J.M. (1963) **Education and Physical Growth**, London, University of London Press.
7. Wilby, P. (1983) **Threat to Single Sex Schools**, London. Sunday Times, 20.11.83.
8. Yates, F. and Pidgeon, D. (1957). **Admission to Grammar School**, Slough, NFER.
9. Murphy, R. J. (1982). Sex differences in objective test performance. **Brit. J. Ed. Psychol.**, 52, 213-219.