/11

APPENDIX F

NATIONAL TESTING AND EQUAL OPPORTUNITIES

Submission by the Equal Opportunities

Commission to the DES Task Group on

National Testing

December 1987

Table of Contents

Summary

1. There are established differences in test performance between girls and boys throughout the age range 7 to 16 and across different areas of the curriculum. Differences in performance at various ages have been shown in language, mathematics, science and verbal and non-verbal reasoning (section 3).

2. Various aspects of test design have differential effects on girls' and boys' performance. The use of practical tests, the form of questions adopted (eg. multiple choice vs. essay questions) and the context and content of individual test items have all been shown to affect the relative performance of girls and boys (3.1.3, 3.1.4, 3.2 and 3.4).

3. These findings have important implications for any proposed national system of testing pupils at 7, 11 14 and 16:

   (i) Where the results of a number of tests are combined to give a single overall test score the resulting gender difference will reflect the weighting given to the component test areas (3.5).

   (ii) If test results are to be used to compare schools, it is important that the proportion of pupils of each sex in the relevant age group within each school should be known and taken into account. The effects of gender differences on inter-school comparison will, of course, be greatest where the comparison involves a single-sex school (4.1).

   (iii) If test results are used to allocate individual pupils to ability groups for particular subjects, differences in test performance will affect the gender composition of different groups. If scores from a number of tests are combined and used for general banding or streaming purposes, the gender balance of each band or stream will be dependent on the weighting given to the different component tests (see 3(i) above) (4.1).

(iv) Tests can be constructed to reflect any desired degree of gender difference, or none. There are no generally accepted objective external criteria for test construction which can guide a test constructor in choosing thc relative weightings for different test questions or topic areas. If tests are constructed, on grounds of equity, to give equal average scores for girls and boys, consistency would require that this should be done at all ages. This would raise questions about other assessment procedures such as GCSE (4.1 and 4.3).

(v) An alternative to eliminating gender differences in average test performance by revising the construction of the tests, is to standardise test scores according to different gender norms. In the EOC's view this would constitute unlawful discrimination (4.2).

Introduction

This report is divided into two parts. The first gives a brief summary and evaluation of the evidence for gender differences and the second part looks at some possible consequences of these in a system of national 'benchmark' tests. The areas studied are those of language, mathematics, ability tests and science.

Because of time limitations, only a very general coverage of the issues is possible. Nevertheless, because the subject is complex, touching upon basic issues of educational achievement, a more extensive treatment is desirable. In a final section an outline is given for such further work.

Evidence for Gender Differences

3.1 Language Differences

3.1.1 Reading

In terms of vocabulary, boys show greater knowledge in Primary school (Douglas et al., 1968) with a persisting difference through the secondary school years (Wittig and Peterson, 1979).

In reading comprehension, an early (7-8 years) advantage in favour of girls, becomes very small by the age of 11 years and by the age of 16 years there is a small advantage to boys (Fogelman et al., 1978; Davie et al., 1972; Douglas et al., 1968; NCDS, 1972).

The American National Assessment of Educational Progress (NAEP, 1986a) reports superior overall performance for girls in reading up to 17 years and the British Assessment of Performance Unit (APU, 1982) does so at ages 11 and 15 years.

3.1.2 Writing

According to NAEP (1986b) girls do better than boys overall at all ages on writing tasks, and this finding is supported by the APU results at 11 and 15 years. NAEP divides writing into 'informative', 'persuasive' and 'imaginative.' In all types of writing girls perform better than boys at all ages from 9 to 17 years.

3.1.3 Test Format

There is considerable evidence that, relatively, boys do better than girls in multiple choice questions when compared to free-response or essay questions. Murphy (1980), dealing with geography O' level exams, finds an increased pass rate for boys when the percentage of multiple choice questions is increased. Wood (1978) reinforces this for language examinations.

3.1.4 Motivation and Context

Wood (1978) provides evidence, based on O' level English language exams, that boys do better than girls when the topic of a question is a, masculine' one (e.g. trains) and girls do better when it is a feminine topic (e.g. a story about a young girl's feelings).

3.2 Mathematics Differences

At the end of primary schooling, boys and girls appear to have similar performances on, largely arithmetical, mathematics tests. By 15 years the boys show an advantage (Douglas et al., 1968). Wood (1976) shows that boys do better than girls in 15 years old public examinations in questions concerning measurement and spatial topics. The APU data (APU, 1986) confirms this at age 15 and also shows that there are only very small differences in number skills and modern algebra. Boys also tend to do better in applied and practical mathematics. At the

age of 11 years boys are more confident in measurement and practical tasks than girls, and at this age the girls performed better than the boys only on computation tasks. There was only a very small difference at this age again in modern algebra. There is also a suggestion that there is only a very small difference at 11 years in problem solving tasks.

Plake et al. (1982), using university students, found that the arrangement of items in a mathematics test was important. With a traditional easy-hard ordering males did better than females but for an arrangement where easy and hard items were uniformly spread across the test these differences disappeared. There appears to be little data concerning such an effect in school age pupils.

3.3 Verbal and Non-Verbal Ability Differences

At the end of primary school, girls appear to score higher than boys on non-verbal tests and considerably higher than boys on verbal tests. By the end of secondary school, boys score somewhat higher on the non-verbal tests. (NCDS 1972; Douglas et al., 1968). Within the non-verbal domain, boys appear to begin to outperform the girls from about age 13 (Macoby and Jacklin, 1974).

3.4 Science Differences

The APU science monitoring programme (APU, 1982) shows that at the age of 11 years boys tend to do better than girls in the application of taught science concepts and in practical investigations. At age 15, results have to be interpreted with care since boys generally have had more exposure to science. The boys do better than girls at this age in the use of equipment, interpreting data, reading information and applying physics concepts. The latter difference is also apparent at age 13.

3.5 Summary

An important feature of gender differences is that context, content and format of test questions can affect these differences, even to the extent of reversing a difference on otherwise similar questions. The multiple choice format appears to favour boys as does practical testing.

4. National 'Benchmark' Testing

It is proposed by the Government that nationally prescribed tests will be administered and marked by teachers at the ages of 7, 11, 14 and 16 years in the areas of mathematics, language and science. It is possible that these ages could be varied somewhat, but that would not affect the substance of what follows. At the time of writing there are no details of how the test scores will be used, but some general indications are available.

The national curriculum consultation document refers to the use of the tests for providing parents with information for comparing schools, and since it is intended that parents will be given details of their children's test results, it is likely that these results will also affect within-school decisions on ability grouping etc. There have been suggestions that the test results will be useful as 'diagnostic' instruments for the teacher and pupils. It is unlikely, however, that the tests will be detailed or sensitive enough for this purpose and most professionals would not accept that a single instrument properly can combine a 'monitoring' and 'diagnostic' function.

The following sections discuss the relevance of gender differences to tests which are designed for purposes of comparison rather than diagnosis.

4.1 The Use of Test Results

It is difficult to predict how the dissemination of test results to parents will affect their actions and perceptions. In terms of local

or central government actions, however, there are several clear possibilities. A LEA could, for example, use test scores as indicators of need and direct resources to schools with poorer performances. Alternatively, it might take the view that poor performing schools should become, say, candidates for closure or amalgamation.

Gender differences will be an important factor where there are disparate proportions of girls across schools, and of course in the case of single sex schools. If the tests produce, for example, lower average scores for girls, then those schools with high proportions of girls will tend to produce lower mean test scores. If such schools 'suffer', either through parent or LEA action then it could be said that gender discrimination was occurring, since relatively more girls than boys would 'suffer' as a result. One way to avoid such an outcome would be to carry out a valid statistical adjustment for gender differences when comparing schools. This will be elaborated upon in the next section.

The other relevant aspect of the use of test results is in individual selection and placement. Thus, a secondary school might have ability groupings for its classes, say for mathematics from the second year. If national mathematics tests scores are available on each child, there would be some pressure, presumably, for these to be used in the ability grouping procedures. In some cases schools may find it difficult to carry out ability groupings in any other way when pupils and parents have access to each child's test results. As pointed out in section 2.5, tests can be constructed generally to reflect any desired degree of gender difference (or none at all) with consequent effects on the numbers of girls and boys in different groupings.

It is generally accepted that there are advantages which follow from belonging to a high ability grouping, so the possibility of discrimination again arises. The problems of constructing gender-fair tests is discussed below. The principal issue here is that there are no generally accepted objective external criteria for test

construction which can guide a test constructor in choosing the relative weightings for different test questions or topic areas. If it were argued, on grounds of equity, that tests should therefore be constructed to give equal average scores for boys and girls, then consistency would require that this should be done at all ages. Furthermore, such a procedure would raise questions about other assessment procedures such as the GCSE and whether attempts should be made to balance results by gender more generally in relevant public exams.

It is worth pointing out that this issue cannot be resolved by appealing to historical precedents in test construction. What has occurred previously partly reflects the culturally related expectations of test constructors concerning gender differences (Gould, 1981).

4.2 Adjusting for Gender Differences

If gender differences are eliminated, in the manner suggested above, or by standardising scores according to separate gender norms, then differences between schools would not be expected to depend on the proportions of girls in the schools. If, however, gender differences in test scores are allowed to persist, then a complicated statistical adjustment would be necessary. A discussion of the problems surrounding such procedures is given by Goldstein (1987).

Suffice it to say here that the use of individual children's test results is required rather than school or LEA averages. It should also be noted that there are other factors, such as ethnic background social class and attainment at time of entry to school, which need to be adjusted for in order to make fair comparisons between schools or LEA's.

It seems that the complexities of the process effectively would rule out its routine use. Certainly it would require a level of expertise unavailable to most LEA's. Moreover, even if such adjustments could

be carried out routinely, similar arguments concerning consistency apply here as in the case of test scores used to compare and group pupils.

4.3 Gender-Fair Tests

It is now commonly accepted good practice that test constructors have an obligation to ensure that their instruments contain no sexual or racial stereotypes or any material which could be offensive to a particular group. There exist guidelines (American Psychological Association et al., 1985), and any national system of testing should aim to follow these broadly. It would also be important that the details of test construction are publicly available so that public scrutiny and challenge is possible if guidelines are not followed.

Of course, even where such guidelines are followed, these cannot be expected to eliminate all gender difference, so that their adoption does not remove the necessity for considering the issues of sections 4.1 and 4.2.

5. Further Work

The review of the evidence of section 3 has not attempted to evaluate existing research findings in terms of their quality. Some studies have been excluded on the grounds of unrepresentativeness, and it is clear that a more thorough review and evaluation of findings, especially concerning narrowly defined domains of achievement, would be useful. Such a review should also pay attention to quantifying differences, so that their relative importance can be assessed.

Some other issues are relevant also. The performance of boys and girls in single and mixed sex schools has been the subject of research, but the existing evidence does not seem to be clear-cut and needs careful evaluation. There have been several studies of teachers' expectations in relation to gender differences and this may well be important if teachers are allowed discretion in the administration and marking of tests. A review of public examination entry policies and results would provide further useful evidence on this issue.

Finally, whatever decisions are made about the implementation of national tests, it is extremely important that there is a proper, and independent, evaluation of their effects with respect to gender as well as other factors. Given the uncertainties about the likely effects, such an evaluation could at least be used to alert people to serious problems and possible violations of equal opportunity legislation.

References

American Psychological Association, American Educational Research Association, National Council on Measurement in Education (1985) Standards for Educational and Psychological Tests. Washington, D.C. A.P.A.

Assessment of Performance Unit (1982) Science in Schools. Survey Reports No. 1 at ages 11, 13, 15. London, DES.

Assessment of Performance Unit (1982) Language Performance in Schools. Survey Reports No.2 at ages 11, 15. London, DES.

Assessment of Performance Unit (1986) A Review of Monitoring in Mathematics, 1978-1982. London, DES.

Davie, R., Butler, N.R. and Goldstein, H. (1982) From Birth to Seven. London, Longman.

Douglas, J.W.B., Ross, J.M. and Simpson, H. (1968) All Our Future. London, Peter Davies.

Fogelman, K., Goldstein, H., Essen, J., and Ghodsian, M. (1978) Patterns of Attainment. Educational Studies, 4, 121-130.

Gould, S.J. (1981) The Mismeasure of Man. New York, W.W. Norton.

Goldstein, H. (1987) Multilevel Models in educational and Social Research. London, Griffin; New York, Oxford University Press.

10. Macoby, E.E. and Jacklin, C.N. (1974) — Psychology of Sex Differences. Stanford, Stanford University Press.

11. Murphy, R. (1980) — Sex differences in GCE Examination entry Statistics and Success Rates. Educational Studies, 6, 169-178.

12. National Assessment of Educational Progress (1986a) — The Reading Report Card. Princeton, Educational Testing Service.

13. National Assessment of educational Progress (1986b) — Writing. Princeton, Educational Testing Service.

14. National Child Development Study (1972) — Second Follow Up of the NCDS. Report to the Social Science Research Council.

15. Plake, B.S., Ansorge, C.J. Parker, C.S. and Lowry, S.R. (1982) — Effects of Item Arrangement, Knowledge of Arrangement, Test Anxiety and Sex on Test Performance. J. Educational Measurement, 19, 49-58.

16. Wittig, S. and Peterson, A. (1974) — Sex Related Differences in Cognitive Functioning. New York, Academic Press.

17. Wood, R. (1976) — Sex Differences in Mathematics Attainment at GCE Ordinary level. Educational Studies, 2, 141-160.

18. Wood, R. (1978) — Sex Differences in Answers to English Language Comprehension Items. Educational Studies, 4, 157-165.

Acknowledgement

This submission is based on a research review undertaken by Professor Harvey Goldstein on behalf of the Equal Opportunities Commission.