What Do We Mean by Equity in Relation to Assessment?

CAROLINE GIPPS
University of London, Institute of Education, 20 Bedford Way, London WC1H 0AL, UK

ABSTRACT *This article reviews the concept of equity in relation to assessment by considering issues of equality of outcome, equal access and the role of context and construct. Performance assessment is discussed as a particular case and evidence is offered from performance assessment in England. The article concludes by discussing approaches to developing assessment practices that are more fair to all the groups likely to be involved in taking those assessments.*

Introduction

Equity and performance assessment are issues currently being debated within the assessment community particularly in the USA. (e.g. Baker & O'Neil, 1994). In this article I want to address what equity in relation to assessment might mean (see also Gipps & Murphy, 1994) and what the implications are for performance assessment.

The UK has a history of using performance assessment even for accountability purposes: the public examination at 16, the General Certificate of Secondary Education (GCSE), involves written responses of the short answer and essay type, practical assessment, oral assessments, and extended pieces of coursework which are assessed in school by the teachers. All marking has a judgmental element and is done centrally by specially trained teachers; it is then moderated either statistically or through inspection. Multiple-choice testing, though it exists in the UK, has never been widely used and is not considered to be an appropriate basis for high-status examinations. At the other end of the age-scale is the National Curriculum Assessment programme for 7 and 11-year-olds; this involves teachers' own assessment of pupil attainment (TA or Teacher Assessment), standardised tests, and some performance-based assessment tasks (STs or Standard Tasks).

Equity is not an underlying theme in education in England and Wales, and indeed there is little clarity about what this might mean. Debate and policy-making where it has featured at all has referred to *equal opportunities* in education with a brief excursion into compensatory education for disadvantaged groups. Early attempts to achieve equality of opportunity, for girls and boys, focused in the main on equality of resources and access to curriculum offerings; important though this is, we now see

0969-594X/95/030271-11

it as a limited approach given the very different out-of-school experiences of girls and boys. The fundamental problem is that this policy focus reflects a deficit model approach to inequality: girls are 'blamed' for behaving like girls and encouraged to behave more like boys. This model implies the possibility of overcoming disadvantage through the acquisition of what is lacking. This approach leaves the status quo essentially unchanged since girls are unlikely to achieve parity through equality of resources and formal equality of access alone. As Yates puts it 'where the criteria of success and the norms of teaching and curriculum are still defined in terms of the already dominant group, that group is always likely to remain one step ahead' (Yates, 1985, p. 212). Equal opportunities is a policy area which has been hotly contested in the UK: it is seen by the extreme right as a revolutionary device which would disturb the 'natural' social order and as an attempt to attack White British society, and by the extreme left as essentially conservative because the gross disparities in wealth, power and status which characterise our society remain unchallenged.

A second approach is one which looks for *equality of outcome* (as evidence of equal opportunities) and this underpins analyses and discussions of group performance at public examination level in the UK. The attitude to equity in the USA is very different from that in the UK, for reasons of history and because of the population structure: 'The US has a long-term commitment to equity for its wholly immigrant population' (Baker & O'Neil, 1994, p. 12) and is evidenced in equal outcome terms: 'The term equity is used principally to describe fair educational access for all students; more recent judicial interpretations, however, have begun the redefinition of equity to move toward the attainment of reasonably equal group outcomes' (Baker & O'Neil, 1994, p. 11) 'the educational equity principle should result in students receiving comparable education yielding comparable performances' (Baker & O'Neil, 1994, p. 12). Our view is that while one must strive to achieve *actual* equality of opportunity, equality of outcomes is not necessarily an appropriate goal: different groups may indeed have different qualities and abilities, and certainly experiences. Manipulating test items and procedures in order to produce equal outcomes may be doing violence to the construct or skill being assessed and camouflaging genuine group differences.

In the UK equal opportunities came to be defined as 'open competition for scarce resources' (Wood, 1987) in the late 1980s and early 1990s. The notion of competition is, however, antithetical to equal outcomes: in a competition the best person wins the prize; competition is *not* designed to offer each individual the best outcome possible for them. In terms of education the latter is, of course, what we seek (while accepting that for some highly selective purposes identifying the 'best' individuals is necessary). Indeed 'fair' competition requires actual equal opportunities and a specification of the rules of the game so that all participants are equally well-prepared.

Apple's (1989) review of public policy in the USA, Britain and Australia leads him to conclude that *equality* has been redefined: it is no longer linked to group oppression and disadvantage but is concerned to ensure individual choice within a 'free market' perception of the educational community. In Apple's view this

TABLE I. Curriculum and assessment questions in relation to equity[a]

Curricular questions	Assessment questions
Whose knowledge is taught?	What knowledge is assessed and equated with achievement?
Why is it taught in a particular way to this particular group?	Are the form, content and mode of assessment appropriate for different groups and individuals?
How do we enable the histories and cultures of people of colour, and of women, to be taught in responsible and responsive ways?	Is this range of cultural knowledge reflected in definitions of achievement? How does cultural knowledge mediate individuals' responses to assessment in ways which alter the construct being assessed?

[a]From Gipps & Murphy (1994) (and after Apple, 1989).

redefinition has reinstated the disadvantage model and underachievement is once again the responsibility of the individual rather than the educational institution.

He argues that attention in the equity and education debate must be refocused on important curricular questions, to which we add assessment questions, in Table I.

To summarise, equity in our view, does not imply equality of outcomes and does not presume identical experiences for all: both of these are unrealistic. Given the contested nature of the equal opportunities concept in the UK, we use instead the term *equity* which is defined in the dictionary as *moral justice*, or *the spirit of justice*. The concept of equity in assessment as we use it implies that assessment practice and interpretation of results are fair and just for all groups. Our focus on equity in relation to assessment considers, therefore, not only the practices of assessment, but also the definition of achievement, whilst at the same time recognising that other factors, e.g. pupil motivation and esteem, teacher behaviour and expectation also come into play in determining achievement.

Equity and Assessment

Where does this debate take us when it comes to assessment? It is important to remember that 'objective' assessment has traditionally been seen as an instrument of equity: the notion of the standardised test as a way of offering impartial assessment is of course a powerful one, though if equality of educational opportunity does not precede the test, then the 'fairness' of this approach is called into question. Most attainment tests and examinations are amenable to coaching and pupils who have very different school experiences are not equally prepared to compete in the same test situation.

As Madaus (1992) points out

> in addressing the equity of alternative assessments in a high-stakes policy-driven exam system policy must be crafted that creates first and foremost

> a level playing field for students and schools. Only then can the claim be made that a national examination system is an equitable technology for making decisions about individuals, schools or districts. (p. 32)

The same point is also made by Baker & O'Neil (1994).

It is important here to consider the role of psychometrics: the traditional psychometric approach to testing operates on the assumption that technical solutions can be found to solve problems of equity with the emphasis on using elaborate techniques to eliminate biased items (Murphy, 1990; Goldstein, 1993). The limitation of this approach is that it does not look at the way in which the subject matter is defined (i.e. the overall domain from which test items are to be chosen), nor at the initial choice of items from the thus-defined pool; nor does it question what counts as achievement. It simply 'tinkers' with an established selection of items. Focusing on bias in tests, and statistical techniques for eliminating 'biased' items, not only confounds the construct being assessed, but has distracted attention from wider equity issues such as actual equality of access to learning, 'biased' curriculum, and inhibiting classroom practices.

Bias in relation to assessment is generally taken to mean that the assessment is unfair to one particular group or another. This rather simple definition, however, belies the complexity of the underlying situation. Differential performance on a test, i.e. where different groups get different score levels, may not be the result of bias in the assessment; it may be due to real differences in performance among groups which may in turn be due to differing access to learning, or it may be due to real differences in the group's attainment in the topic under consideration. The question of whether a test is biased or whether the group in question has a different underlying level of attainment is actually extremely difficult to answer. Wood (1987) describes these different factors as the opportunity to acquire talent (access issues) and the opportunity to show talent to good effect (fairness in the assessment).

When the existence of group differences in average performance on tests is taken to mean that the tests are biased, the assumption is that one group is *not* inherently less able than the other. However, the two groups may well have been subject to different environmental experiences or unequal access to the curriculum. This difference will be reflected in average test scores, but a test that reflects such unequal opportunity in its scores is not strictly speaking biased, though its use could be invalid.

Hence, to achieve equity in assessment, interpretations of students' performance should be set in the explicit context of what is or is not being valued: an explicit account of the constructs being assessed and of the criteria for assessment will at least make the perspective and values of the test developer open to teachers and pupils. A considerable amount of effort over the years has gone into exploring cognitive deficits in girls in order to explain their poor performance on science tests; it was not until relatively recently that the question was asked whether the reliance on tasks and apparatus associated with middle-class white males could possibly have something to do with it. As Goldstein (1993) points out, tests are framed by the test developers' construct of the subject and their expectations of differential performance.

Of course pupils do not come to school with identical experiences and they do not have identical experiences at school. We cannot, therefore, expect assessments to have the same meaning for all pupils. What we must aim for, though, is an equitable approach where the concerns, contexts and approaches of one group do not dominate. This, however, is by no means a simple task; for example, test developers are told that they should avoid any context which may be more familiar to males than females or to the dominant culture. But there are problems inherent in trying to *remove* context effects by doing away with passages that advantage males or females, because it reduces the amount of assessment material available. De-contextualised assessment is anyway not possible, and complex reasoning processes require drawing on complex domain knowledge. Again, clear explanation of the constructs is important.

In an assessment which looks for best rather than typical performance, the context of the item should be the one which allows the pupil to perform well, but this suggests different tasks for different groups which is in itself hugely problematic. However, what we can seek is the use, within any assessment programme, of a range of assessment tasks involving a variety of contexts, a range of modes within the assessment, and a range of response format and style. This broadening of approach is most likely to offer pupils alternative opportunities to demonstrate achievement if they are disadvantaged by any one particular assessment in the programme.

Indeed, this is included in the Criteria for Evaluation of Student Assessment Systems by the National Forum on Assessment (NFA):

> —to ensure fairness, students should have multiple opportunities to meet standards and should be able to meet them in different ways
> —assessment information should be accompanied by information about access to the curriculum and about opportunities to meet the Standards
> —...assessment results should be one part of a system of multiple indicators of the quality of education. (NFA, 1992, p. 32)

Performance Assessment

Performance assessments (PA) are assessments which model real tasks, i.e. require the pupil to perform in the assessment what we wish them to learn in the classroom; usually they focus on higher levels of cognitive complexity; require the production of a response (in a range of modes); and require qualitative judgements to be used in the marking. PA is an important development in educational assessment with implications for equity, so I will devote a section to it.

The difference in the UK and USA settings in relation to PA should not be underestimated: as explained in the Introduction, assessment in the UK is predominantly on the PA model; within PA we can list a range of approaches from the high status written examination, the Standard Tests and Tasks of the National Curriculum assessment programme and teacher-assessed coursework in GCSE, to portfolios and Records of Achievement (RoAs). As PA is rather less-well developed in the US, there is a tendency to use the generic term: hence the concern among minority

groups that 'alternative equals non-standard equals sub-standard' (Baker & O'Neil, 1994). In the UK, not all PAs are considered equal: there is a world of difference between the public examination and the Record of Achievement and this is reflected in the status of these assessments: a Record of Achievement would not be considered sufficiently external and rigorous for selection and accountability purposes. The amount of information provided is also an issue: a percentage mark, grade or level is easier to use (for anything other than teaching purposes) than is the more qualitative, descriptive information from RoAs.

It is important, therefore, to distinguish between say portfolio assessment and specified PA tasks which are set for trained teachers and marked by them, against specified criteria using agreed marking systems, with the system underpinned by moderation. These differences in approach will have significant effects on consistency of approach and scoring, affecting the construct assessed and inter-rater reliability, both of which are highly pertinent to equity issues, particularly if the assessments are used for high-stakes purposes.

Furthermore, PA cannot be developed using traditional psychometric techniques for analysing items, because far fewer items are involved and there is no assumed underlying score distribution. This may force a shift towards other ways of reviewing and evaluating items based on qualitative approaches, for example sensitivity review, a consideration of the construct, how the task might interact with experience, etc.; such a development is to be welcomed.

The question which seems to be being addressed in the US is, is PA a good form of assessment? This question has, however, to be deconstructed into: a good form of assessment for what purpose? and better than which other forms of assessment? Our experience in the UK would suggest that high stakes PA can change curriculum focus and broaden teaching: this happened as a result of the introduction of the GCSE at 16 (HMI, 1988) and at age 7 with National Curriculum assessment (Gipps *et al.*, 1995). The strain on teachers which this sort of change brings should not be underestimated, however, nor indeed their need for in-service training in the subject area and the new assessment model.

For school-based assessment and formative assessment PA is a better model than standardised tests—because of its flexibility and potential to assess constructs in more depth. Furthermore, it is possible to use a highly structured, externally marked and moderated PA programme for accountability and certification purposes; the resources required to support such a programme are significant, however, and the English experience is that it is manageable at only one or two points in the system (Gipps, 1995) (indeed there should be no need for more than this).

The introduction of The English National Curriculum Assessment Programme, which was based on a PA model, raised and illuminated some important equity issues; the programme has changed over the years but the lessons remain of value.

Performance Assessment and Equity Issues in National Curriculum Assessment

The National Assessment programme requires that pupils are assessed across the

full range of the National Curriculum using external tests and teachers' own assessment. The external tests were originally called Standard Assessment Tasks (SATs) and the teachers' assessments are called Teacher Assessment (TA). The original SATs used in 1991 and 1992 were true performance assessments and involved classroom-based, externally-set, teacher assessed activities and tasks. For example, at age 7, reading was assessed by the child reading aloud a short passage from a children's book chosen from a list of popular titles, using dice to play maths games, using objects to sort, etc. At age 14 there were extended investigative projects in maths and science and assessments based on classroom work in English. What they had in common, across both ages, as well as the performance element was classroom, rather than formal examination, administration; interaction between teacher and pupil with ample opportunity to explain the task; and a reduced emphasis on written responses—particularly at age seven.

The role of communication in PA, involving spoken and written responses together with understanding of the instructions for the task, can present a significant threat to equity for minority language groups.

Evidence from the piloting of the SATs for 7-year-olds in 1990 indicated that the bilingual children seemed more insecure initially when presented with new work in the SATs; when this was the case, the peer group became a very important source of support. In an assessment situation, however, this posed difficulties for the teacher in deciding whether the intervention of another pupil had clarified the child's understanding of the question or supplied the correct answer. The misunderstanding of instructions was a serious problem for bilingual pupils: they appeared to relax and respond better when questions were rephrased in the mother tongue; they became more motivated and handled tasks more confidently. When activities were lengthy and complex there was a particular burden on bilingual children and examples of misunderstanding did not always come to light. Teachers felt that the bilingual children found it particularly difficult to show their true ability in maths and science. This was largely due to the difficulty of assessing oral responses in science interviews and the difficulties these children experienced in the group discussion element of science and maths investigations (NFER/BGC, 1991).

The teachers also reported a hazard in small group testing: where children worked in mixed groups for assessment, the boys were sometimes more dominant and girls took a passive role, a commonly observed pattern of gendered performance.

In our study of national assessment in primary schools [1], teachers reported strong feelings that the national assessment programme at age seven was inevitably unjust for bilingual children (Gipps, 1992). Since their English language skills were still in the early stages of development, these children were disadvantaged in any assessment that was carried out for comparative purposes. These teachers felt that formal summative (or accountability) assessment for comparison is, at this age, unfair for such children and thus runs counter to their notion of equity. These teachers had similar views about children from disadvantaged backgrounds but their feelings about bilingual learners were particularly strong.

That said, there was a feeling, however, that the SATs, for all their difficulties of classroom management, time and unmanageability, and despite their heavy reliance

on language, did offer a better assessment opportunity for children with special needs and bilingual learners, than would more traditional 'paper and pencil' tests. Our teachers' views were that, whatever the level of disadvantage for their bilingual learners in summer 1991—and this was where their anxieties lay, not with gender issues—that this would be increased in the more formal testing which they anticipated in summer 1992. Recent research has confirmed this (Clarke & Gipps, 1995).

Furthermore, children who were second language learners tended to perform less well than other pupils on the SATs, but there was some evidence that they performed better on the SATs than in the TA, and this was a fairly widespread finding. This suggests that structured PA was fairer for minority pupils than (unmoderated) TA (and, one could deduce, better than nothing) since in effect, teacher stereotyping was being challenged.

In the piloting of SATs for 14-year-olds in 1991 there was a detailed investigation of teachers' views in relation to these assessment tasks in maths (CATS, 1991) but not in other subjects. The teachers administering the SATs felt that the nature of the SAT rendered it accessible to pupils who were not fluent in English. Aspects which contributed to this included: interaction with the teacher, the practical elements of the tasks, a normal classroom atmosphere, interactions with other pupils and the variety of presentation and assessment modes. The conclusion made was that for pupils who are not fluent in English, written materials cannot enable the demonstration of attainment without teacher–pupil interaction. Most of these teachers felt that pupils who were not fluent in English could engage in the SAT activity. Thus the style of the activity was appropriate for most of these pupils. However, only a third of teachers thought that the SAT enabled pupils to demonstrate appropriate attainment. This comment no doubt is related to the fact that overall the attainment of non-fluent pupils was below that of others in both the SAT and the TA. However, analysis of the SATs showed that pupils who were *not* fluent were scoring higher on the SAT than in the TA which suggests that the TA awarded by the teachers may have been an underestimate due to the pupils' perceived language difficulties, and that the SATs facilitated high performance for non-fluent pupils to a greater extent than it did for others. The pilot report of the maths scheme states that: 'if pupils who are not fluent in English are to be entitled to a fair assessment it is essential that the SATs retain the interactive, practical and flexible aspects' (CATS, 1991).

The report from the Schools Examination and Assessment Council (SEAC) which drew together the KS3 findings in the 1991 pilot (SEAC, 1991) from the various agencies did not discuss performance by ethnic group (presumably because the sample sizes were small) but pointed out that for bilingual learners performance was 'relatively high'. This they felt was because teachers were able to provide the normal classroom support for these pupils during SATs, and the materials were generally accessible to pupils whose home language was not English.

Unfortunately, more recent developments in the national assessment programme have not been subjected to the same level of piloting and analysis, but an important point emerges from these early National Curriculum Assessment Studies. The SAT-type activity with its emphasis on active, multi-mode assessment and detailed

interaction between teacher and pupil may, despite the heavy reliance on language, be a better opportunity for minority and special needs children to demonstrate what they know and can do than traditional, formal tests with limited instructions. The key aspects seem to be:

- a range of activities;
- match to classroom practice;
- extended interaction between pupil and teacher to explain the task;
- normal classroom setting;
- a range of response modes other than written.

Furthermore, many of these pupils performed better on the SAT than in the TA and this made the teachers think hard about their evaluation of the pupils.

Conclusion

There is no such thing as a fair test, nor could there be: the situation is too complex and the notion simplistic. However, by paying attention to what we know about factors in assessment, administration and scoring, we can begin to work towards tests that are more fair to all the groups likely to be taking them, and this is particularly important for assessment used for summative and accountability purposes. On the grounds of equity, all groups must be offered actual equality of access to the curriculum, and examinations and assessment must be made as fair as possible to all groups.

So how do we ensure that assessment practice and interpretation of results is as fair as possible for all groups? It is likely that a wide ranging review of syllabus content, teacher attitude to boys and girls and minority ethnic groups, assessment mode and item format is required, as Table I shows, if we wish to make assessment as fair as possible. Although this is a major task, it is one which must be addressed in the developing context of national standards, national curriculum, and national assessment. An example that it is possible and what is required is given by the case of physics Higher School Certificate in South Australia: girls' performance in physics has improved dramatically since 1988 when a female chief examiner was appointed. This examiner has deliberately worked within a particular model of physics (which takes a 'whole view' of the subject); simplified the language of the questions; included contexts only that are integral to particular physics problems; offered a range of different ways of answering questions which does not privilege one form of answer over another; provided a list of key instruction words and examples of how students would go about answering questions which include these words (ESSSA, 1992).

We need to encourage clearer articulation of the test/exam developers' construct on which the assessment is based, so that the construct validity may be examined by test takers and users. Test developers need to give a justification for inclusion of context and types of response mode in relation to the evidence we have about how this interacts with group differences and curriculum experience. The ethics of

assessment demand that the constructs and assessment criteria are made available to pupils and teachers, and that a range of tasks and assessments be included in an assessment programme. These requirements are consonant with enhancing construct validity in any case. Given the detailed and, as yet, poorly understood effect of context (Murphy, 1993) on performance, the evidence that girls more than boys attend to context in an assessment task, and the ability of changes in the context of the task to alter the construct being assessed, the area of construct validity demands detailed study. We certainly need to define the context of an assessment task and the underlying constructs and make sure they reflect what is taught.

We must encourage the use of a range of modes and task style; we need also to expand the range of indicators used:

> Multiple indicators are essential so that those who are disadvantaged on one assessment have an opportunity to offer alternative evidence of their expertise. (Linn, 1992, p. 44)

Assessment which elicits an individual's best performance involves tasks that are concrete and within the experience of the pupil (an equal access issue) presented clearly (the pupil must understand what is required of her if she is to perform well) relevant to the current concerns of the pupil (to engender motivation and engagement) and in conditions that are not threatening (to reduce stress and enhance performance) (after Nuttall, 1987).

Although we do not look for equality of outcome, we must continue to seek genuine equality of access; this means that all courses, subjects studied, examinations, etc. are actually equally available to all groups *and* are presented in such a way that all groups feel able to participate fully. One suggestion from the United States is that, since opportunity to learn is a key factor in performance, schools may have to 'certify delivery standards' as part of a system for monitoring instructional experiences (Linn, 1993). How realistic it is to do this remains to be seen, but it does put the onus on schools to address the issue of equal access, at an actual rather than formal level.

We also need to be clear about what counts as proper preparation of pupils in any assessment programme. If there are preparation practices which are considered to be unethical then they should be spelled out. The other side of the coin is that teachers and schools have a commitment to teach pupils the material on which they are going to be formally assessed. To this requirement we should add proper preparation of teachers so that they understand the basic issues in assessment and are equipped to carry out good formative assessment.

A fundamental question in equity and assessment is: are group differences in measures 'real' or are they the result of the measuring system? The answer, of course, is likely to be 'a bit of both' and what we need to do is to minimise the latter, while understanding and articulating causes of the former. I hope that this article and the one by Patricia Murphy in the same issue go some way towards developing our understanding of these complex issues.

Notes

Note that an earlier version of this paper was presented to the AERA conference, New Orleans, 1994.

[1] *National Assessment in Primary School: an evaluation*, ESRC Grant No. R 000 23 21 92.

References

APPLE, M.W. (1989) How equality has been redefined in the conservative restoration, in: W. SECADA (Ed.) *Equity and Education* (New York, The Falmer Press).

BAKER, E. & O'NEIL, H. (1994) Performance Assessment and Equity: a view from the USA, *Assessment in Education*, 1, pp. 11–26.

CATS (1991) *Pilot Report 1991 Key Stage 3 Maths* (London, SEAC).

CLARKE, S. & GIPPS, C. (1995) *The Impact of National Assessment at Key Stages One & Two*, Report to the National Union of Teachers, Institute of Education, London.

ESSSA (1992) *Gender Equity in Senior Secondary School Assessment Project*, Third Progress Report September 1992, Senior Secondary Assessment Board of South Australia.

GIPPS, C. (1992) National testing at seven: what can it tell us?, paper presented at *AERA Conference* 1992, San Francisco.

GIPPS, C. (1994) *Beyond Testing: towards a theory of educational assessment* (London, The Falmer Press).

GIPPS, C. (1995) Reliability validity and manageability in large scale performance assessment, paper presented at *AERA Conference*, Atlanta, 1993, in: H. TORRANCE (Ed.), *Evaluating Authentic Assessment* (Milton Keynes, Open University Press).

GIPPS, C. & MURPHY, P. (1994) *A Fair Test? Assessment, achievement and equity* (Milton Keynes, Open University Press).

GIPPS, C., BROWN, M., MCCALLUM, B. & MCALISTER, S. (1995) *Intuition or Evidence? Teachers and National Assessment of Seven Year Olds* (Milton Keynes, Open University Press).

GOLDSTEIN, H. (1993) Assessing group differences, *Oxford Review of Education*, 19, pp. 141–150.

HMI (1988) *The Introduction of the GCSE in Schools 1986–1988* (London, HMSO).

LINN, M.C. (1992) Gender differences in educational achievement, in: *Sex Equity in Educational Opportunity, Achievement and Testing* (Princeton, NJ, ETS).

LINN, R.L. (1993) Educational assessment: expanded expectations and challenges, *Educational Evaluation and Policy Analysis*, 15, pp. 1–16.

MADAUS, G. (1992) A technological and historical consideration of equity issues associated with proposals to change the nation's testing policy, paper presented at *Symposium on Equity and Educational Testing and Assessment*, March, Washington, DC.

MURPHY, P. (1990) Gender differences—implications for assessment and curriculum planning, paper presented at *BERA Conference*, Roehampton, August.

MURPHY, P. (1993) Some teacher dilemmas in practising authentic assessment, paper presented to *AERA Annual Meeting*, Atlanta.

NATIONAL FORUM ON ASSESSMENT (1992) Criteria for evaluation of student assessment systems, *Educational Measurement: issues and practice*, Spring, p. 32.

NFER/BGC (1991) *An Evaluation of National Curriculum Assessment*, Report 3, June.

NUTTALL, D. (1987) The validity of assessments, *European Journal of Psychology of Education*, 11, pp. 109–118.

SEAC (1991) *National Curriculum Assessment at Key Stage 3: a review of the 1991 pilots with implications for 1992* (London, EMU: SEAC).

WOOD, R. (1987) Assessment and equal opportunities, text of public lecture at ULIE, 11 November.

YATES, L. (1985) Is 'girl friendly schooling' really what girls need?, in: J. WHYTE, R. DEEM, L. KANT & M. CRUIKSHANK (Eds) *Girl Friendly Schooling* (London, Methuen).