Chapter 6
Fairness in Assessment

Caroline Gipps and Gordon Stobart

Introduction

Fairness is a concept for which definitions are important, since it is often interpreted in too narrow and technical a way. We set fairness within a social context and look at what this means in relation to different groups and cultures. Similarly, we are using *educational assessment* in a more inclusive way than is often the case; we include tests, examinations, teachers' judgments or evaluations ('assessment' in the United Kingdom) of student performance. We then explore *bias* in measurement and how it relates to validity, as well as the broader concept of *equity*. Finally, three examples of approaches to ensure fairness are given.

We argue that 21st-century assessment will need to take ever more account of the social contexts of assessment and to continue the movement away from seeing fairness simply as a technical concern with test construction. Fairness in assessment involves both what precedes an assessment (for example, access and resources) and its consequences (for example, interpretations of results and impact) as well as aspects of the assessment design itself.

Fairness

How would we tell whether a test is fair for different groups (male/female; socially/advantaged/disadvantaged; ethnic groupings)? The dilemma is that different groups will have different qualities and experiences, so fairness in assessment cannot be judged in terms of equal scores or outcomes.

Differences in performance on a test may be due to differing access to learning, or because the test is biased in favour of one group. Wood (1987) described these different aspects of fairness as the opportunity to acquire talent (access issues) and the opportunity to show talent to good effect (fairness in the assessment).

In our view, fairness in assessment cannot be considered in isolation from access issues in the curriculum and the educational opportunities offered to the students:

C. Gipps (✉)
Vice-Chancellor, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY
e-mail: c.gipps@wlv.ac.uk

C. Wyatt-Smith, J.J. Cumming (eds.), *Educational Assessment in the 21st Century*,
DOI 10.1007/978-1-4020-9964-9_6,

fairness in access opportunities both to schooling and to the curriculum provide the 'level playing field' that must precede a genuinely fair assessment situation.

Fairness and Equity

We use the term 'equity' interchangeably with 'fairness'. *Equity* is defined in the *Chambers Concise Dictionary* (1992) as 'moral justice'. Equity does not imply equality of outcome and does not presume identical experiences for all—both of these are seen to be unrealistic, but it asserts that assessment practice and interpretation of results need to be fair and just for all groups.

For example, it is possible to have similar outcomes for two groups and yet to see this as unfair to one of them, which may have been disadvantaged in terms of access to the curriculum. Conversely, it is possible to have unequal group outcomes that may be seen as fair. An example would be where there are group differences in the application to learning and preparation, where each had similar resources and opportunities.

Equity is also a quasi-legal term. The legal meaning of *equity* is 'the spirit of justice' and, building on the work of Walter Secada (1989), we see it as a qualitative concern for what is just. 'Equity attempts to look at the justice of a given state of affairs, a justice that goes beyond acting in agreed upon ways and seeks to look at the justice of the arrangements leading up to and resulting from those actions' (p. 81).

The implication is that equity is not the same as equality. Equity represents the judgment about whether equality, be it in the form of opportunity and/or of outcomes, achieves just ('fair') results. Looking for equality requires essentially a quantitative approach to differences between groups, while equity goes beyond this and looks at the justice of the arrangements prior to the assessment.

The approach we take includes these broader issues and, therefore, owes more to sociocultural theory than to measurement theory. Sociocultural research and theory builds on Vygotsky's work, in which it is used as a specific term embodying the roles of social interaction and cultural context in learning and identity formation (Cobb, 1994; Penuel & Wertsch, 1995.) Although assessment is a key player in the learner's formation of identity (Gipps, 1999), we do not focus on that aspect of sociocultural approaches to assessment in this chapter. Rather, we take a view of assessment that places it in social, cultural and political contexts: assessment is a socially embedded activity that can only be fully understood by taking account of the social and cultural contexts within which it operates, alongside the technical characteristics.

A Brief History of Assessment and 'Fairness'

There is a significant history of assessment being used for fairness and equity purposes. This stems from the belief that testing is fairer than selection by patronage or birth, since all sit the same test under the same conditions. This, as we shall show, is a very restricted view of fairness.

Selection

Selection has probably been the most pervasive role of assessment over the years (Glaser & Silver, 1994). Assessment for selection, which later became linked with certification, illustrates well the power and control aspects of assessment as well as its role in cultural and social reproduction.

Examinations were first developed in China under the Han dynasty (206 BC to AD 220) in order to select candidates for government service. The Jesuits introduced competitive examinations into their schools in the 17th century, possibly influenced by Jesuit travellers' experience in China. It was not until the late 18th century and early 19th century that examinations developed in northern Europe—in Prussia and then in France and England—again, in order to select candidates for government positions.

In Europe, as the industrial capitalist economy flourished, there was an increasing need for trained middle-class workers. Access to the professions had been determined, before the 19th century, by family history and patronage rather than by academic achievement or ability. In the 19th century this picture began to change. The economy required more individuals in the professions and in managerial positions. Society, therefore, needed to encourage a wider range of individuals to take on these roles. This was the first time that upward mobility became a practical proposition on a wide scale. Of course, there had to be some way of selecting those who were deemed suitable for training, as well as certifying those who were deemed to be competent, and examinations were used as the tool. The appeal of examinations was that they were the same for everyone who took them, though, of course, this was generally restricted to educated males. Thus, although the exams limited nepotism and corruption, they could not eliminate the advantages afforded by gender, social status and wealth. In Britain, in the case of the Civil Service exams, for example, it was still almost exclusively those who had received an appropriate fee-paying education who were able to pass.

Assessment for selection has also been a key theme *within* school systems. In the United Kingdom and elsewhere, intelligence testing has historically played a central role both in identifying those considered able enough for an academic secondary education and selecting out of the system those with special educational needs deemed more suitable for 'special' schools, an approach enshrined in the 1944 *Education Act* in the United Kingdom (Sutherland, 1996). The validity of intelligence (IQ) tests as a fair means of selection has come under increasing scrutiny (Gardner & Cowan, 2005). It is now widely recognised that IQ tests are culturally based and biased in favour of individuals from the dominant culture. Therefore, the sociocultural critique of intelligence testing is that it obscures the perpetuation of social inequalities because it legitimates them (Gould, 1996; Hanson, 1993).

Equity was also a driving force behind the development of 'objective' tests. By 'objective' we are referring to multiple-choice tests and others that require no judgment in scoring. From their post-World War I origins onwards, the development of objective tests for sorting and selecting students was seen, particularly in the United States, as a scientific, even progressive, activity (Stoskopf, 2008; Ryan, 2008). The growth of such testing has grown exponentially in the United States (Madaus &

Raczek, 1996) and its efficiency as a method of mass assessment has increasing appeal around the world. Such tests have highly replicable and reliable scoring—hence the 'objective' label. This appeal has often obscured the limited validity of such tests and the subjective nature of item writing, selection of material and formulation of answer choices.

Of longer pedigree is the more open-ended ('constructed response') tradition of written examinations, though the critique is in many ways similar (Broadfoot, 1979). There may be added concern that examinations, with their demands for culturally dependent forms of response (for example, the argumentative essay), may penalise those from more disadvantaged or culturally different backgrounds as there may be a mismatch between the language and culture of the home and the school. As a result, examinations may offer a less-than-fair assessment, and furthermore, because of their role in certification, they may institutionalise and legitimate social stratification (Stobart, 2008).

To summarise, although external examinations, IQ testing and objective testing were seen originally as equitable tools for selection and certification purposes, a sociocultural critique calls this into question. Assessment, in its various forms, has a determining role to play in cultural reproduction and social stratification. The discussion of fairness in this chapter needs, therefore, to be set against this background.

Developments in Assessment and Their Relationship to 'Fairness'

There have been considerable developments in the nature and conceptualisation of assessment over the past 50 years. These have often been the result of the changing purposes for which assessment has been used. One example is the use of testing for accountability purposes, particularly the use of targets based on the results of high-stakes testing such as the *No Child Left Behind* testing program in the United States (Stobart, 2008). Such programs raise the issue of fairness in large-scale testing, which we address later in the chapter.

The second example (which has received increasing emphasis) is the use of assessment to contribute to the learning process, in general terms called 'educational assessment'. It is to fairness issues in this approach that we now turn.

The Move to Educational Assessment

Building on the critiques of IQ testing, and developments in understanding of how learning takes place, researchers—mostly in the United States at first—began to conceptualise different types of, and approaches to, assessment, usually with an educational purpose rather than an 'organisational' one such as selection.

In the development of educational assessment, the work of Glaser was critical. His 1963 article on criterion-referenced testing was a watershed in the development

of a new type of assessment, which moved away from classical testing based on psychometric theory. Glaser (1963) made the point that norm-referenced testing developed from psychometric work that focused on aptitude, selection and prediction. Educational assessment, by contrast, aimed to devise tests that look at the individual as an individual, rather than in relation to other individuals, and to use measurement to identify strengths and weaknesses individuals might have, so as to aid their educational progress. The development of this criterion-based approach, rather than one based on norms, was not driven by fairness but can be seen as a fairer approach.

New developments—performance assessment, 'authentic' assessment, portfolio assessment and so forth—were part of a move to design assessment that supports learning and provides more detailed information about students (Wolf, Bixby, Glenn, & Gardner, 1991). We can see this, also, as a shift towards 'an opening up' of traditional assessment, an approach that can itself be seen as a fairness issue.

However, focus on an assessment approach on its own is not sufficient for a discussion of fairness. Consideration must still be given to students' opportunity to learn (Linn, 1993), the knowledge and language demands of the task (Baker & O'Neil, 1994) and the criteria used for scoring (Linn, Baker, & Dunbar, 1991). Clearly, as with traditional forms of assessment, questions of fairness arise in the selection of tasks and in the grading of responses. Furthermore, the more informal and open-ended such assessment becomes, the greater the reliance on the judgment of the teacher/assessor. The strength of classroom assessments is that a broader range can be assessed than in a timed examination, increasing validity, while reliability may benefit from repeated assessments. A threat to reliability, however, may come from any bias in the teacher's judgment, either in the form of negative stereotyping or a 'halo' effect for favoured students. These may themselves reflect cultural attitudes about, for example, gender.

What we do know is that a broadening of assessment approaches will offer students alternative opportunities to demonstrate achievement if they are disadvantaged by any one particular assessment in a classroom or program. According to Linn (1992), '[m]ultiple indicators are essential so that those who are disadvantaged on one assessment have an opportunity to offer alternative evidence of their expertise' (p. 44).

Fairness and Validity

Our claim that fairness should be seen within a sociocultural frame rather than as a technical exercise mirrors, a shift that has taken place in developments around the concept of validity. In this section we claim that fairness should be embedded within validity arguments rather than treated as a separate and often 'add-on' concept. This is because current validity theorising incorporates concerns about fairness and bias, and reflects similar understandings of the social basis of assessment.

At the heart of the reformulation of validity is the move from treating it as a *fixed property* of an assessment to seeing it as *process* that investigates an assessment

in terms of both the construct being assessed (how effectively it sampled the target domain) and, crucially, the inferences and actions based on the results. The 1999 United States *Standards for educational and psychological testing*[1] takes this approach:

> Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests. (p. 9)

The importance of this for considerations of fairness and bias is that we cannot declare a test to be unfair or biased until we know what the purpose of the testing was and how the results were interpreted. Our argument that fairness is a sociocultural issue, rather than simply a technical one, is the same as the argument advanced for this understanding of validity. Validity is not simply the way in which a test functions, but depends on what it is used for and the interpretation and social consequences of the results. This was recognised by Samuel Messick in his seminal 1989 chapter:

> For a fully unified view of validity, it must also be recognised that the appropriateness, meaningfulness and usefulness of score-based inferences depend as well on the social consequences of the testing. Therefore, social values cannot be ignored in considerations of validity. (p. 19)

Incorporating Fairness Concerns into Validity Arguments

An essential part of validity is the concern with whether the inferences made from the results of an assessment are fair to all those who were assessed. If a test has sampled a domain in a way that benefits a particular group, then its validity is reduced, since the inferences drawn from the results may be misleading. As we have already seen, this is the error of assuming that a test is 'fair' because candidates sat the same test at the same time—without consideration of whether some candidates were privileged in terms of preparation for it. This may then be further compounded by the privileged candidates' interpretation of their performance in terms of merit and natural ability, so that their success can then be put down to merit rather than privilege—a Victorian line of reasoning that is still with us today (Stobart, 2008). Equity concerns about what precedes an assessment are therefore a part of the validation of the assessment. Validity enquiry must also involve construct validity and the interpretation and consequences of the results.

We provide three examples of validity enquiries that focus on fairness: large-scale assessments, test construction and teachers' assessments of their students.

---

[1] AERA, APA, & NCME (1999). *Standards for educational and psychological testing*. Washington, D.C.: AERA. Michael Kane's definitive chapter on validity in the 4th edition of *Educational measurement* (2006) takes a similar approach.

Example 1: Fairness in Large-Scale Multicultural Assessments

This example emphasises the role of construct validity by looking at the assumptions made about what is assessed. We take the position that there is no cultural neutrality in assessment or in the selection of what is to be assessed, and attempts to portray any assessment as 'acultural' are a mistake. Cumming (2000) observes that 'Acultural knowledge has definite cultural roots. This is knowledge that is privileged in our standards and testing procedures' (p. 4). She goes on to raise two key questions, which link with those in Table 6.1:

1. When setting standards and test content, are we really sure this is the knowledge we need?
2. Are we really privileging certain knowledges to maintain a dominant culture and in doing so ensuring perpetuation of ourselves, as people who have succeeded in the formal educational culture to date?

These concerns are central to fairness and validity. This line of reasoning has implications both for what is sampled in an assessment and how we interpret the results if we know some groups have been disadvantaged in both access and preparation. These are summarised in Table 6.1.

In every country, there will be examples of groups being disadvantaged in terms of access and preparation. For example, Meier (2000) has reported that in South Africa the teacher–learner ratio was 1:40 for black learners compared to 1:21 for whites. This was compounded by a shortage of qualified teachers in mathematics and science, which meant that many schools for black students did not even offer these subjects, even though they were part of the official curriculum. Mwachihi and Mbithi (2000) reported how in Kenya the introduction of 'cost sharing' has

Table 6.1 Access, curriculum and assessment questions in relation to equity and validity

Access questions	Curricular questions	Assessment questions
Who gets taught and by whom?	Whose knowledge is taught?	What knowledge is assessed and equated with achievement?
Are there differences in the resources available for different groups?	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?
What is incorporated from the cultures of those attending?	How do we enable the histories and cultures of people of colour, and of women, to be taught in responsible and responsive ways? (Apple, 1989)	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? (Gipps & Murphy, 1994)

Source: Stobart, 2005.

meant that schools now have to fund the purchase of books and other materials, leaving schools in poorer areas without adequate resources. This has been exacerbated by the introduction of a more complex, centrally devised curriculum that is deemed irrelevant to regional needs. In the United States, inequalities in access and preparation have been addressed through highly controversial 'affirmative action' approaches in which disadvantaged, but lower-scoring, students were given priority. This has been increasingly subject to legal challenge. This has been mirrored in England by prestigious universities such as Bristol offering admission to students in state schools in preference to some students from private schools, who may have had similar or better grades. There has been a considerable media backlash, stoked by parents who have paid for their children's education and who now see themselves as disadvantaged. In China, the disadvantages for rural minority groups have been recognised by setting differential pass standards on its Higher Education Entrance Examination (Zhao, 2000).

These examples illustrate how the validity concerns about how the results are interpreted meld with fairness concerns about what has gone on before the assessment itself and how results should be interpreted and acted upon.

Example 2: Fairness in Test Development

Equity concerns with access and preparation overlap with test development, even though fairness in test development has often been reduced to statistical consideration of bias in test items. Our argument is that simply seeking to minimise item bias is insufficient; tests take place in a social context and this needs consideration.

However, seeking to create tests that are as fair as possible to different groups is a necessary part of the process. The risk is that it may lead to a concern with presentational features rather than with which constructs are being sampled and how. This restricted view of bias is captured in the fairness section of *Standards for educational and psychological testing* (AERA, APA, & NCME, 1999):

> A full consideration of fairness would explore the many functions of testing in relation to its many goals, including the broad goal of achieving equality of opportunity in our society . . . The *Standards* cannot hope to deal adequately with all these broad issues . . . Rather the focus of the *Standard*s is on those aspects of tests, testing and test use that are the customary responsibilities of those who make, use and interpret tests. (p. 73)

This is also reflected in the six Educational Testing Service (ETS) International Principles for Fairness Review of Assessments (2004):

Principle 1.	Treat people with respect in test materials.
Principle-2	Minimise the effects of construct-irrelevant knowledge or skills.
Principle-3	Avoid material that is unnecessarily controversial, inflammatory, offensive, or upsetting.
Principle-4	Use appropriate terminology to refer to people.
Principle-5	Avoid stereotypes.
Principle-6	Represent diversity in depictions of people.

Source: ETS, 2004.

In relation to validity, these are seeking to avoid what Messick (1989) called 'construct irrelevant variance', features that are likely to interfere with the assessment of a construct; in this case by distracting or upsetting a candidate or drawing on something culturally unfamiliar. We now look at some of the issues that have to be addressed within this more restricted approach to bias.

Test Bias

A test is biased if 'two individuals with equal ability (in the subject being tested) but from different groups do not have the same probability of success' (Shepard, Camilli, & Averill, 1981).

A cause of bias in a test could be that it was designed by one cultural group to reflect their own experience, and thus disadvantages test takers from other cultural groups, an accusation levelled at IQ tests. Thus, bias may be due to the content matter in a test, or lack of clarity in instructions, which leads to differential responses from different groups. Bias may also be due to scoring systems that do not credit appropriate or correct responses that are more typical of one group than the other.

Gould (1996) provides us with an extreme historical example of questions asked of newly arrived non-English speaking immigrants:

> Crisco is: patent medicine, disinfectant, toothpaste, food product;
> Christy Mathewson is famous as a: writer, artist, baseball player, comedian.

They also had to respond to verbal instructions such as:

> When I say 'go' make a figure 1 in the space which is in the circle but not in the triangle or square, and also make a figure 2 in the space which is in the triangle and circle but not in the square. Go.
>
> (Gould, 1996, p. 230)

If we wish students to do well in tests/exams, we need to think about assessment that elicits an individual's best performance (after Nuttall, 1987). This may involve tasks that are concrete and within the experience of the student (an equal-access issue), presented clearly (the student must understand what is required of them if they are to perform well), relevant to the current concerns of the student (to engender motivation and engagement), and in conditions that are not threatening (to reduce stress and enhance performance) (Gipps, 1994).

We are now well aware that the form of assessment can differentially affect results for different groups. In England, there has been far more analysis of this in relation to gender than to ethnicity. We know that during compulsory schooling (up to 16 years) girls are likely to outperform boys on tasks that involve open-ended writing, particularly when this involves personal response. Even within multiple-choice tests, traditionally seen as favouring boys, there are differential response patterns. In the United States, Carlton (2000) has shown that in such tests, females perform better than males, matched for ability, on questions in which the content is a narrative or is in a humanities field and when the content deals with human relationships. As the context of an item grows longer the relative performance of females also improves. Males outperform females on questions relating to science,

technical matters, sports, war or diplomacy. We also know that where examinations have a coursework (or essay) element, the performance of girls is likely to be more consistent, though the effect this has on final grades in English school-leaving exams has often been overstated (Elwood, 1995).

We know less about other aspects of the form of assessment, particularly in relation to ethnicity. For example, oral assessment plays little part in the examination system in England outside examining languages. Does the emphasis on written response disadvantage groups who place more emphasis on oral communication in their culture?

The existence of group differences in average performance on tests is often taken to imply that the tests are biased, the assumption being that one group is not inherently less able than the other. However, as we have argued, the two groups may well have been subject to different environmental experiences or unequal access to the curriculum. This difference will be reflected in group test scores, but the test is not, strictly speaking, biased.

One of the key statistical measures for identifying potential item bias in multiple-choice tests is the use of differential item functioning (DIF):

> A statistical measure related to fairness should be used, whenever sample sizes permit, as an empirical check on the fairness of questions. Statistical measures based on the way matched people in different groups perform on each test question, called differential item functioning or DIF, are preferred. DIF occurs when people in different groups perform in substantially different ways on a test question, even though they have very similar scores on the test. If DIF data are available, tests should be assembled following rules that keep DIF low.
>
> (ETS, 2004, p. 11)

While the intention with DIF is laudable, we have reservations about how this may undermine construct validity. The requirement should be to select assessment content *that accurately reflects the construct*, even if it produces gender/ethnic group differences, and to avoid content that is not relevant to the construct and that could affect such differences. This again takes us beyond a technical exercise to broader considerations in which different interests need to be recognised. It should also be noted there is nothing equivalent to DIF to guide construction of other forms of assessment, apart from professional judgment and examination of overall grades for different groups.

Example 3: The Fairness and Validity of Teachers' Informal Evaluations/Assessments

Fairness in assessment in the informal setting of the classroom can be both more difficult—because there are many complex issues for the teacher to consider—and more possible, since a range of assessment approaches is possible. It is more feasible for the teacher to offer, in the informal assessment setting, a range of assessment tasks and modes, an approach that supports fairness as we argued above. It is also more feasible to provide the situation that can elicit an individual's best performance, since it is under the teacher's control.

Referring back to our introductory espousal of a sociocultural stance, a crucial aspect of this approach to assessment includes allowing students the tools to help them show what they can do, and arguably the most important tool is the teacher. In classroom-based assessment, there is opportunity for teacher and students to clarify/discuss the objective being assessed, how it might be assessed and what counts as success or mastery. Such an approach brings the student into a more active role in the learning process and helps to build self-evaluation and meta-cognitive skills and is thus good learning practice (Black & Wiliam, 2006; Edwards, 2005; Pryor & Crossouard, 2008). Through this, students from a range of backgrounds also have the chance to have their strengths and understandings recognised. This undeniably places demands on the teacher, and staff development may be required to ensure that the teacher is open to such new interpretations and, indeed, relationships. Thus, the developing corpus of work on sociocultural approaches to assessment has implications for fairness, although these implications have not been explicitly addressed.

However, teachers' informal assessment is, to a certain extent justifiably, perceived as being unreliable and biased (Harlen, 2004). This is often to do with lack of clarity, and variability, in standards or criteria. It is possible to improve the consistency of teachers' assessments through: providing clear criteria, training teachers to assess against these, and supporting the process with moderation of judgments via discussion (ARG, 2006).

It is also possible that teachers' cultural values could lead to bias in the assessment. These may themselves reflect cultural attitudes about, for example, gender, with research showing that in the United Kingdom noisy young boys are more likely to be marked down by teachers (Harlen, 2004). Baker & O'Neil (1994) also showed how the use of portfolios, regarded by their advocates as a progressive move towards authentic assessment, were viewed by some minority groups in the United States as a white, middle-class activity which disadvantaged those with fewer resources and opportunities.

In relation to the curriculum offered and opportunity to learn, there is another inconvenient fact: teacher expectation can affect the curriculum and learning experiences offered to children. There is clear evidence that teachers offer a different curriculum to children for whom they hold low and high expectations (Harlen, 2004; Tizard, Blatchford, Burke, Farquhar, & Plewis, 1988; Troman, 1988). This is pertinent to the equal-access issue.

Conclusion

Fairness is both essential and elusive. It is the appeal to fairness that has made educational 'measurement' a pivotal part of most cultures. We have argued that different groups being allowed to sit, and be judged by, the same test is a simplistic view. Fairness needs to be linked to equality of opportunity, which includes access to similar resources and curricular opportunities. The more familiar, and narrower, discussion of bias in testing is only a small part of this.

The challenge for 21st-century assessment is to broaden our views of fairness to take fuller account of social and cultural contexts. The temptation, however, is to back away from the larger social issues because they are difficult, and to concentrate on the assessment itself, for example, in relation to bias. Just as the theorising of validity has moved from it being a property of a test to a process based on how the results are interpreted, we can envisage a move to the discussion of fairness focusing on the inferences made about the results and the impact of these. So we move away from talking about a biased test to talking about interpreting the results in a way that is fair to all the groups taking the assessment. The debates around positive discrimination and allowing for disadvantage would be a part of this.

We will never achieve fair assessment, but we can make it fairer: The best defence against inequitable assessment is openness. Openness about design, constructs and scoring and grading will bring out into the open the values and biases of the test design process, offer an opportunity for debate about cultural and social influences and open up the relationship between assessor and learner. These developments are possible, but they do require political will.

References

AERA, APA, & NCME (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education). (1999). *Standards for educational and psychological testing*. American Educational Research Association: Washington, DC: AERA.

Apple, M. W. (1989). How equality has been redefined in the Conservative restoration. In W. Secada (Ed.), *Equity and Education*. New York: Falmer Press.

Assessment Reform Group (ARG). (2006). *The role of teachers in the assessment of learning*. Assessment Reform Group pamphlet. See also ARG website at <www.assessment-reform-group-org>.

Baker, E., & O'Neil, H. (1994). Performance assessment and equity: A view from the USA. *Assessment in Education, 1*(1), 11–26.

Black, P., & Wiliam, D. (2006). Developing a theory of formative assessment. In J. Gardner (Ed.), *Assessment and learning* (pp. 81–100). London: Sage.

Broadfoot, P. (1979). *Assessment, schools and society*. London: Methuen.

Carlton, S. T. (2000). *Contextual factors in group differences in assessment*. Paper presented at 26th IAEA Conference, Jerusalem.

Cobb, P. (1994). Where is the mind? *Educational Researcher, 23*(7), 13–20.

Cumming, J. (2000). *After DIF, What culture remains?* 26th IAEA Conference, Jerusalem.

Edwards, A. (2005). Let's get beyond community and practice: The many meanings of learning by participating. *The Curriculum Journal, 16*(1), 49–65.

Elwood, J. (1995). Undermining gender stereotypes: Examination and coursework performance in the UK at 16. *Assessment in Education, 2*(3), 283–303.

Educational Testing Service (ETS). (2004). *ETS International principles for fairness review of assessments*. Princeton NJ: Author.

Gardner, J., & Cowan, P. (2005). The fallibility of high stakes '11-Plus' testing in Northern Ireland. *Assessment in Education: Principles, Policy and Practice, 12*, 145–165.

Gipps, C. (1994). *Beyond testing, towards a theory of educational assessment*. London: Falmer Press.

Gipps, C. (1999). Sociocultural aspects of assessment. *Review of Research in Education, 24*, 357–392.

Gipps, C., & Murphy, P. (1994). *A fair test? Assessment, achievement and equity*. Buckingham: Open University Press.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18*, 519–521.

Glaser, R., & Silver, E. (1994). Assessment, testing, and instruction: Retrospect and prospect. In L. Darling-Hammond (Ed.), *Review of research in education* (pp. 393–419). Washington, DC: American Educational Research Association.

Gould, S. J. (1996). *The mismeasure of man*. New York: Norton.

Hanson, F. A. (1993). *Testing: Social consequences of the examined life*. Berkeley: University of California Press.

Harlen, W. (2004). *A systematic review of the evidence of reliability and validity of assessment teachers use for summative purposes*. In *Research evidence in education library* Issue 3, London: EPPI-Centre, Social Science Research Unit, Institute of Education. Retrieved April 4, 2007, from http://eppi.ioe.ac.uk/EPPIWeb/home.aspx?page=/reel/review_groups/assessment/review_three.htm.

Linn, M. C. (1992). Gender differences in educational achievement. In J. Pfleiderer (Ed.), *Sex equity in educational opportunity, achievement and testing*. Princeton, NJ: Educational Testing Service.

Linn, R. L. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis, 15*, 1.

Linn, R. L., Baker, E., & Dunbar, S. (1991). Complex, performance based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15–21.

Madaus, G. F., & Raczek, A. E. (1996). The extent and growth of educational testing in the United States: 1956–1994. In H. Goldstein (Ed.), *Assessment: Problems, developments and statistical issues*. Chichester: John Wiley & Sons.

Meier, C. (2000). The influence of educational opportunities on assessment results in a multicultural South Africa, Paper presented at the 26th IAEA Conference, Jerusalem.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: American Council on Education and Macmillan.

Mwachihi, J. M., & Mbithi, M. J. 2000. Assessment and equity assurance in the Kenyan multicultural background, Paper presented at 26th IAEA Conference, Jerusalem.

Nuttall, D. (1987). The validity of assessments. *European Journal of Psychology of Education, 11*(2), 109–118.

Penuel, W., & Wertsch, J. (1995). Vygotsky and identity formation: A sociocultural approach. *Educational Psychologist, 30*(2), 83–92.

Pryor, J., & Crossouard, B. (2008). A socio-cultural theorisation of formative assessment. *Oxford Review of Education 2007, 34*(1), 1–20.

Ryan, A. M. (2008). *From child study to efficiency: The use of testing in the Chicago public schools, 1899 to 1928*. Paper presented at the American Research Association's Annual Meeting, New York.

Secada, W. G. (1989). Educational equity versus equality of education: An alternative conception. In W. G. Secada (Ed.), *Equity and education*. New York: Falmer Press.

Shepard, L., Camilli, G., & Averill, M. (1981). Comparison of procedures for detecting test item bias with both internal and external ability criteria. *Journal of Educational Statistics, 6*, 317–375.

Stobart, G. (2008). *Testing times: The uses and abuses of assessment*. London & New York: Routledge.

Stobart, G. (2005). Fairness in multicultural assessment systems. *Assessment in Education, 12*(3), 275–287.

Stoskopf, A. (2008). Sowing grain and cultivating roses: IQ testing and educational reform in the Boston public schools, 1910–1932. Paper presented at the American Research Association's Annual Meeting, New York.

Sutherland, G. (1996). Assessment: Some historical perspectives. In H. Goldstein (Ed.), *Assessment: Problems, developments and statistical issues*. Chichester: John Wiley.

Tizard, B., Blatchford, P., Burke, J., Farquhar, C., & Plewis, I. (1988). *Young children at school in the inner city*. Hove: Lawrence Erlbaum Associates.

Troman, G. (1988). Getting it right: Selection and setting in a 9–13 middle school. *British Journal of Sociology of Education, 9*(4), 402–422.

Wolf, D., Bixby, J., Glenn, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. *Review of Research in Education, 17*, 31–74.

Wood, R. (1987). *Measurement and assessment in education and psychology*. London: Falmer Press.

Zhao, H. (2000). *The minority nationality related issues in China public examinations*. Paper presented at the 26th IAEA Conference, Jerusalem.