Towards Equity in Assessment: crafting gender-fair assessment

BAGELE CHILISA
Department of Educational Foundations, University of Botswana, Post Bag 0022, Gaborone, Botswana

ABSTRACT *This article explores definitions of achievement and their relationship to gender-fair assessment. A framework to discuss achievement is drawn from the affinities of standpoint theory and post-structural theories' emphasis on the role of language in transmitting norms and values that shape gender relations. The first proposition put forward is that different modes of assessment denote different forms of knowledge that are not necessarily gender-neutral. The article illustrates the argument by identifying patterns of gender-biased items in the Junior Certificate science examinations in Botswana. The conclusion drawn is that these patterns illustrate forms of knowledge that are representative of male and female ways of perceiving reality. The patterns of gender-biased items corroborate those found in the international literature. The second proposition is that assessment materials may be gender-biased. The article illustrates gender bias through a content analysis of the Junior Certificate English reading passage examination and discusses the implications of this example for fairer assessment. Finally, a checklist is suggested for crafting gender-fair assessment.*

Introduction

Assessment, especially when it takes the form of a national examination, is the most powerful tool that those who control the schools use to assert their power. In a male-dominated patriarchal society, examinations may accentuate and perpetuate gender inequalities in academic achievement by controlling and manipulating what counts as knowledge. Whenever gender inequalities in academic achievement are observed, we ought to ask the following questions: How is achievement defined? Who defines achievement and in whose interest? What is the purpose of achievement? Who grades? Who defines the criteria for grading? What messages do the language, the content and materials used in the assessment tasks convey? This article will explore definitions of achievement and their relationship to gender. Beyond these substantive issues, the article will address issues related to the crafting of gender-fair assessment.

A framework to discuss achievement in relation to equity is drawn from the affinities of standpoint theory (Harding, 1986) and post-structural theories on the role of language in transmitting norms and values that shape gender relations. The

perspective of the author is that the definition of achievement is heavily dependent on what counts as knowledge. Knowledge, on the other hand, may be defined in terms of a subject or discipline. Within a subject, the type of ability required for one to achieve, the context within which one is expected to achieve and the task formats, content and learning targets emphasised by the curriculum may all be perceived as forms of knowledge. Within this perspective, the article starts with a review of the perceived differences in achievement between males and females in the assumed forms of knowledge, and then considers the implication for fairness in assessment. It then explores the perceived forms of knowledge that emerge from an analysis of multiple-choice items in the national Junior Certificate science paper written in Botswana in 1995.

The second proposition put forward in the article is that assessment materials convey certain values and that assessment may be biased if the task materials and the language are demeaning to one gender. The impact of such assessment may or may not be revealed in the differences in performance between the sexes. It does, however, illustrate the way an education system seeks to socialise males and females and influence the way boys and girls perceive themselves. The Junior Certificate English examination reading passage, written in Botswana in 1995, is used to demonstrate forms of gender bias that have implications for fairness in assessment.

Perspectives on Assessment and Forms of Knowledge

The Standpoint Theory and Subject Choice

The argument of standpoint theory is that knowledge is always referenced to some standpoint (Thompson & Gitlin, 1995). What counts as knowledge is tied to the interests and perceived purposes of knowledge of different interest groups. Definitions and the construction of knowledge are therefore political. Conventional knowledge, with its various mainstreams in the form of subjects, derives its legitimacy from the claims, assumptions and values of the dominant male group. Thus, standpoint theory provides a framework for debating and contesting what counts as relevant, important and significant knowledge. From the standpoint theory perspective, the superior performance of boys may depend on devalued performance criteria for girls.

Can assessment tools be gender-neutral? Neutrality implies fairness. Fairness can be judged by the forms of knowledge which are assessed and equated with achievement. The most pertinent question is whether the range of cultural knowledge that is gender-related is reflected in definitions of achievement (Gipps & Murphy, 1994). The knowledge to be assessed can be defined by the subject. For instance, science is often regarded as a 'male' subject because the definition of knowledge in science stems from male Western philosophers and as such the traits valued in the discipline are those which may be considered male-oriented, such as rationality, objectivity, activity-orientation, selfishness and competitiveness. Science subjects like physics and chemistry are more popular with boy students. Meanwhile, females may have opposite traits, less valued in the science discipline, such as emotionality,

subjectivity and nature-directness. Female traits may also include beauty-valuing, passivity, selflessness and co-operativeness (Hildebrand, 1996). Girls therefore tend to prefer humanities subjects such as languages, literature and history.

This subject choice is not so much an outcome of biological determinism. It is to a large extent a result of the socialisation process. It is a reflection of the persistent and enduring cultural forces that sort and slot students into different subjects. In Botswana, which is a predominantly patriarchal society, day-to-day social life is regulated and sanctioned by cultural norms that reinforce differences between males and females. In pre-independence Botswana, children were socialised through a formal education system undertaken in initiation schools called *bojale* for girls and *bogwera* for boys (Shapera, 1959; Townsend-Coles, 1985; Moorad, 1993). The education was based on the expected gender roles, behaviours and responsibilities, with *bogwera* emphasising knowledge of inheritance, rearing of cattle, land and hunting, which are science-related. *Bojale* emphasised domestic roles, such as caring for the young and the sick, obedience and co-operation, building huts, cooking and other domestic chores. Today, initiation ceremonies are practised in only two districts in Botswana but stereotyped gender role expectations persist. For some ethnic groups, the formal socialisation of girls is still undertaken during ceremonies marking menarche. At these ceremonies the girl is instructed on the behaviours that mainly reinforce the superior status of men and the subordinate status of women.

These differences manifest themselves in marked disparities in participation in science in Botswana. At the senior secondary level, where there is a subject choice, the ratio of boys to girls in science in 1998 was 3:1. At university level, in the 1993–94 academic year, 76% of the fourth year B.Sc. students were males (Taiwo & Molobe, 1994). These disparities have been linked to attitudes among other factors. Taiwo & Molobe (1994) showed that boys aged 16–19 preferred science-related subjects such as mathematics, design and technology and woodwork, while girls preferred language and arts subjects and disliked science-related subjects, especially physics. They contend that stereotyping based on the socialisation process may possibly account for these dichotomised views.

One of the ways to achieve gender fairness in assessment is to seek a balance in the number of boy-oriented and girl-oriented subjects that count towards achievement. Maintaining such a balance may be an essential strategy to ensure that no knowledge is devalued.

Gendered Asymmetry in the 'Measuring Stick of Worth'

While the standpoint theory lays the ground for contesting existing forms of knowledge, post-modernism argues for the rejection of knowledge as absolute truth. It critiques claims to objectivity and rationality and posits that knowledge should be negotiated with the participants. To date, the criteria for success in education have been dominated by positivism, with its emphasis on rationality and objectivity. Positivism has implications for the relationship between the learner and what is assessed, and for the observer judging the learning. Learners adopt a passive role because what is learnt resides outside their realm. It is pre-determined and assumes

TABLE I.

Male trait	Female trait
Abstract	Holistic
Quantitative	Qualitative
Outcomes	Process
Competition	Co-operation
Objective	Subjective
Knower/mind	Knowable/nature
Hierarchical	Multiplicity
Value-free	Value-laden

Source: Hildebrand (1996), pp. 151–152.

an absolute truth. Assessment tools under these circumstances require students to reproduce in undigested form what they learn from the teachers. Objective tests like multiple-choice test items, matching type test items and fill-in-the-blanks type test items are perceived as the best techniques for gathering data on student achievement.

Alternative forms of assessment, mostly qualitative in nature, such as performance assessment tasks and portfolio tasks, are undervalued for the reasons that the scoring is subjective and time-consuming and the scores are unreliable. Qualitative assessment tools and techniques are holistic, integrative and subjective in nature, requiring assessment of those practices linked to student life experiences. In these forms of assessment, the students have a role in negotiating what is to be assessed and how it is assessed. Teacher evaluation of students is complemented by student self-evaluation and peer evaluation. Group work is valued and assessed. These alternative assessment techniques are more complementary to the females' ways of knowing.

There is a recognised gendered asymmetrical dualism in the 'measuring stick of worth', where concepts inclined towards traits valued by women are devalued (Hildebrand, 1996, p. 151). Hildebrand categorises this dualism as shown in Table I.

When any one of these traits is dominant in an assessment task, a response mode or the types of ability demanded by the task, then a certain definition of knowledge is implied. For example, subjective/objective task formats, and process versus product learning outcomes, may be considered as based on distinctive forms of knowledge. Within an assessment mode, one can view item features, content examined and context of the assessment task as representing different forms of knowledge. To achieve fairness, the choice of knowledge to be used as evidence of learning should therefore seek a balance between those inclined towards the males' ways of knowing and those inclined towards the females' ways of knowing.

Forms of Knowledge and Assessment Modes

Task Formats: Subjective/Objective

Knowledge in assessment may be defined by task formats such as paper-and-pencil task formats, performance formats, long-term activity formats and personal com-

munication formats. There is evidence of gender asymmetry in relation to task formats. For instance, boys do better than girls in the test tasks of the multiple-choice type (Ben-Shakar & Sinai, 1991). However, in most developing countries, the Primary School Leaving Examination (PSLE) has been dominated by the multiple-choice test format. Yet the PSLE is a selection examination, selecting only up to 50% of the candidates to continue into secondary school. The dominance of one task format in such a significant examination may be discriminating against girls.

Process versus Product

Knowledge may also be defined according to the learning targets emphasised by the curriculum. For instance, demonstration of process may be the targeted learning outcome or it may be the product that is targeted (Nitko, 1996). The evidence suggests that females are better at process skills while males are more likely to excel when the assessment targets a product (Stobart *et al.*, 1992). Continuous assessment may be viewed as evaluating the process of learning while national examinations are summative and emphasise product. The weighting of continuous assessment and national examination is therefore an important aspect in judging fairness in assessment.

Types of Ability

Knowledge may also be defined according to the type of ability demanded by the assessment task. Studies suggest that males out-perform females in general information ability, arithmetical reasoning and spatial ability, while females out-score males in verbal ability, spelling, grammar, language usage, rote memory and perceptual speed (Maccoby & Jacklin, 1974; Kimura, 1992). Males tend to have the ability to extract spatial and logical relationships independently of the contextual components of the tasks (Levy, 1980). Females tend to explain the meaning of a concept from connotative content, as shown, for example, in the background context. They are also able to form associations between seemingly unrelated ideas, a pattern of reasoning usually neglected in science (Shemesh, 1990). Also, males' ways of knowing may be characterised by an emphasis on competition, learning in the abstract and hierarchical thinking (Rennie & Parker, 1991), while females' ways of knowing are more inclined towards co-operation, learning in context and holistic thinking (Tobias, 1990). The gender-differentiated variation in types of ability suggests the need to shift from using behaviourist schemes of structuring knowledge, such as Bloom's taxonomy (Bloom *et al.*, 1956), to forms constructed on the basis of qualitative approaches to structuring knowledge, such as the qualitative approach of Perry (1970). Such a shift in the types of knowledge valued by assessment would produce greater fairness.

Content Form and Out-of-school Experience

Knowledge may also be defined according to the type of out-of-school learning experiences that form the context of the assessment task. Gender bias in context and

TABLE II.

Male	Female
Public domain	Domestic domain
Culture	Nature
Status role	Relational to man

format is revealed when questions asked influence how students respond, even when knowledge of the context is not required to solve the problem. For instance, boys might do well on maths questions set in a sports context while girls might do well in maths questions involving human relations. Female students may do better in maths problems that involve using a formula, while male students may do better when the route to solving the quantitative problem is unclear or when estimation provides shortcuts to solutions. Thus to avoid gender bias, assessment tasks should reflect both females' and males' out-of-school experiences.

Post-structural Theories, Language and Assessment

Bias in assessment may also be revealed in the values and prejudices embodied in the materials and the use of language that may interact with assessment tasks to produce differences in performance. Post-structural theories analyse the pattern of human culture and lived experience as well as language. They underscore the power of language and discourse in seeking to understand experiences. Language is regarded as a cultural symbol and by far the most important transmitter of our values and prejudices. Every lexical verb not only communicates the action to be taken but also has an underlying message about our beliefs, values and prejudices. Language affects how people think and behave (Henley, 1989).

Frameworks for reviewing gender bias in language can borrow from stereotypical general differences between men and women that have been conceptualised in terms of sets of metaphorical binary opposites (Ortner & Whitehead, 1991). The procedure for content analysis looks at contexts depicted by language use that define men in status and role categories as warriors, statesmen and chairmen in contrast with women, who are defined in terms of their relationship to men. Other oppositions include men's alignment with culture and women's with nature (Levi-Strauss, 1969; Ortner, 1972) and men's alignment with the public domain and women's with the domestic domain (Rosaldo, 1974). These categories fall into an asymmetrical dualism illustrated in Table II.

Content analysis for bias in language can also look at contexts where language ignores women through the use of the masculine as a generic form, for example in the words 'manpower' and 'chairman'. Other contexts include the use of the generic term when referring to one gender, for instance mentioning parents when in fact

reference is being made to mothers. Subtle language bias arises when one gender is consistently named before the other and when such a characteristic even acquires a grammatical rule (Eichler, 1991). The question is whether gender-biased language affects student performance. In the context of assessment, research has shown that language affects the cognitive processing of assessment tasks. Silveira (1978) found that reaction time from stimulus to correct response was longer for subjects responding to a task where the masculine generic was applied to a female picture than where it was applied to a male picture. Crawford & English (1984), in a study of memory, found that females' recall of essays 48 hours later was better when they were written in unbiased form than when they were written in generic terms. This research serves to show that language may alienate the learner from the subject learnt, producing in the learner unconscious forms of resistance that delay processing of the assessment task.

On the basis of the above arguments, it can be concluded that the various dimensions of assessment modes, such as the subject, the content form, the task formats and the outcome targeted, can all be viewed as embodying particular forms of knowledge. Examination materials may also contain gender-biased language that may affect the performance of boys and girls in the assessment task. This is illustrated in the following case-studies. The first reveals gendered patterns in item responses in a multiple-choice item national examination using the Mantel–Haenszel (MH) χ^2 procedure. The second reveals gender-biased language in a reading passage used in a national examination using a content analysis procedure.

Case-study 1: the Botswana Junior Certificate Science Multiple-Choice Examination, 1995

Method

A sample of 42 community junior secondary schools (CJSSs) was selected from 180 CJSSs in Botswana. A trend analysis of each school's performance in the years 1989–1994 was made first. Schools in the upper quartile, middle quartile and lower quartile were identified. Schools in the three quartiles were then randomly selected from a group of schools in the urban, semi-urban and rural areas. Data on students' performance on the science multiple-choice examination were obtained from a total of 4265 female respondents and 3999 male respondents from the 42 schools. The age of the respondents ranged between 14 and 16 years.

The MH χ^2 procedure was used to identify items that show differences in responses between males and females. This statistical analysis provided a technique to produce empirical evidence that made visible the differences in examination item responses between males and females. A statistical analysis of the item responses in the Botswana Junior Certificate examination of 1995 shows how, in perceiving gender as a cultural schema, it is possible to identify patterns that go beyond individual idiosyncrasies. The examination items were also treated as texts that were interrogated and deconstructed in order to seek meaning and understanding of the

possible processes and practices that result in the dichotomies. The items identified were regarded as artefacts coming from male and female cultural schemes.

A Review of the Technique

Various statistical procedures can be used to identify items that show differences in responses between males and females. Objective types of items could be analysed for item difficulty, where item difficulty is simply the percentage of the respondents in each group getting the item correct. The difference between male and female item performance, referred to as item impact (Holland & Thayer, 1988), has been used to identify items that are differentially difficult for males and females (Erickson & Erickson, 1984; Bateson & Parsons-Chatman, 1989). It has been argued that item difficulty is an inadequate measure of differential item functioning because the index is confounded with other characteristics of the item, such as discrimination power (Angoff, 1982; Holland & Thayer, 1988). For instance, difference in proportion pass for two groups on the more discriminating items is greater than the proportion pass on the less discriminating items even when the items are of the same difficulty level. The method may therefore identify as bias items that are highly discriminating (Angoff, 1982). Current methods avoid the simple comparison of proportion pass and compare proportion pass on the item among examinees who are equal in the ability and skill being measured. This implies that the procedure sorts examinees according to performance in the test. The examinees are then grouped according to their scores in the test. Examinees with similar scores are then treated as if they have the same ability and skills. Among the popular approaches used are methods based on item response theory (Hambleton & Rogers, 1989) and χ^2 procedures (Holland & Thayer, 1988). Although the three-parameter item response theory approach is the theoretically preferred procedure (Shepard *et al.*, 1985), the MH χ^2 procedure is often used because it includes significance tests and is less costly (Holland & Thayer, 1988).

Samples of test item responses for males and females could be analysed to detect differential item functioning (DIF) using the MH method. The statistical procedure looks for items on which one group out-performs another in spite of apparently similar knowledge. The procedure shows the proportion pass (p value) for each group, and the impact value. It also shows the MH χ^2 (MH-CHSQ), which tests the hypothesis that the δ values are equal to zero. The MH-CHSQ flags an item as exhibiting DIF if, within a group of examinees with scores in the same test score level, the proportion of examinees responding correctly to the same item is not the same for the subpopulation groups. It uses the idea of observed–expected cases for each cell frequency at each score level just like any regular χ^2 test. It is a non-parametric measure distributed as a χ^2 with 1 degree of freedom. The δ value is a p value converted to a normal deviate with an arbitrarily chosen mean and standard deviation (Angoff, 1982). The procedure further shows a weighted odds ratio that compares the odds of answering the item correctly for members of the reference group (males in this study) with the odds of answering the item correctly for members of the focal group (females in this study). Items with a weighted odds ratio

TABLE III. ETS DIF classification scheme

Category	Amount of DIF	Value of δ
A	Non-significant	Less than 1.0 or not significantly different from 0 at $\alpha = 0.05$
B	Moderate	Greater than or equal to 1.0 and significantly different from 0 at $\alpha = 0.05$
C	Heavy	Greater than or equal to 1.5 and significantly greater than 1.0 at $\alpha = 0.05$

greater than 1.0 are differentially easier for males while items with a value of less than 1.0 are differentially easier for females. When the two groups have the same odds of answering the item correctly, the value of the ratio is 1.0, indicating that there is no differential item functioning. A transformation of the odd ratio value by multiplying it by -2.35 puts it into the same scale as differences in δ values (MH D-DIF). MH D-DIF is thus an effect measure of the amount of DIF. It is preferable to the MH-CHSQ because it takes into consideration the difficulty level of the item. The transformed statistics have a value of zero when there is no DIF in an item, while negative and positive values indicate the direction of the DIF. In this study negative values indicate items that are differentially easier for the males while positive values indicate items easier for the females.

The MH procedure further shows an item category that flags items depending on the absolute value of the δ statistic for the item and its statistical significance. The DIF category is a classification of items by degree of DIF used by the Educational Testing Service (ETS) (Dorans, 1989; Zwick & Erickan, 1989). Table III summarises the ETS classification scheme.

The MH procedure flags an item as operating differently for population subgroups but does not indicate the causes of differences. A qualitative review of the DIF items is then used to identify and explain patterns in the gendered items. What follows is a discussion of the outcome of the analysis.

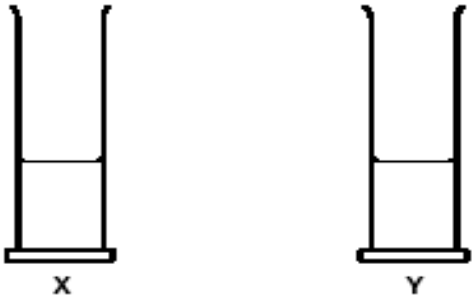
Illustrative Examples of Gendered Patterns of Responses to Items

A statistical analysis of the 1995 science Junior Certificate national examination revealed differences in performance between boys and girls patterned around perceived utility of content, item features and context (see also Chilisa, 1997; Chilisa & Matambo, 1997).

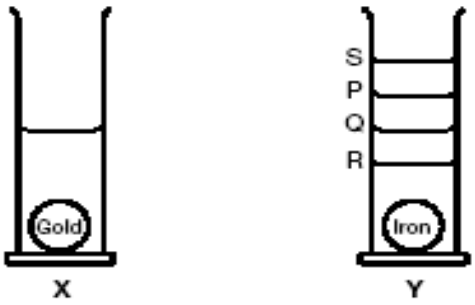
Boys' Performance. Out of 60 items, 11 were biased. Of these, six favoured the girls while five favoured the boys. Three characteristics distinguish these items: item features, content areas and context used in the task. Three of the five items which favoured boys, 5, 11, and 37, were presented in diagram form (see Fig. 1).

Item 5 was presented in diagram form and required measurement of volume.

Mpho has two balls of the same size. One ball is made of gold and the other made of iron. He took two measuring cylinders **X** and **Y** containing the same volume of water.



He put the ball made of gold in measuring cylinder **X** and the water level rose to mark Q. He then put the ball made of iron in cylinder **Y**.

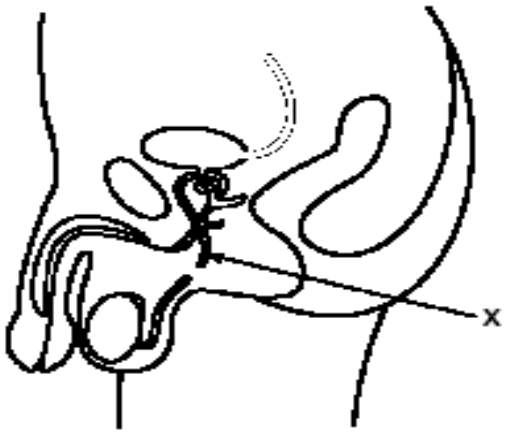


5. To which mark did the water level rise in measuring cylinder **Y**?

A. P
B. Q
C. R
D. S

Contrary to the assertion that diagrams assist thinking, it has been found that they may be of hindrance to girls, who have limited spatial ability compared to boys (Orton, 1987). The diagram may partly explain the large DIF effect size: 151.61. The large effect size may also be the result of a direct transfer of knowledge of units of measurement in mathematics, where the greatest difference between males and females was observed (Orton, 1987), to science. In addition, in Botswana, boys have been shown to conserve volume earlier than girls, which may also explain their superior performance on the item on volume.

Item 11 also had a significant DIF effect size: 280.95. The item requires identification of a male sexual organ shown in an unnamed diagram. The females



Study the diagram above and use it to answer question 11.

11. The part labelled **X** is the

A. Sperm duct
B. Ureter
C. Urethra
D. Fallopian tube

37. Which of the four diagrams below would be easiest to lift the load?

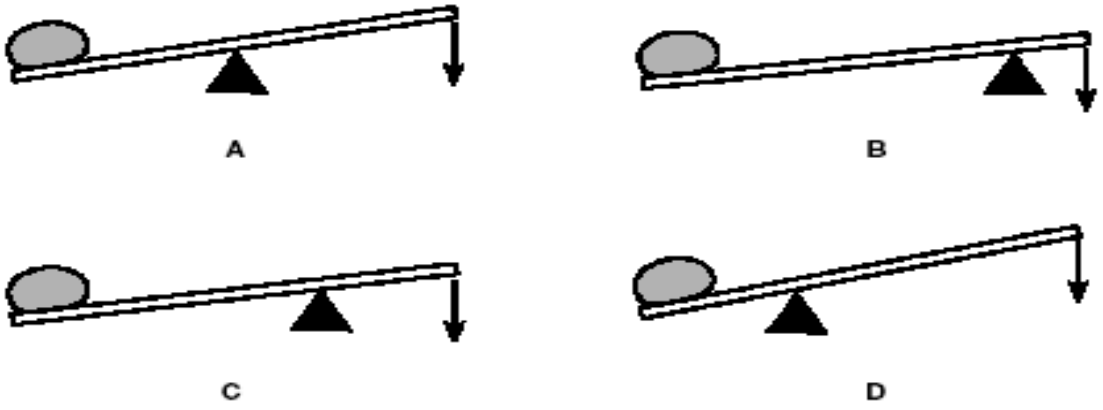


FIG. 1. Diagram items.

would have to struggle to find out what the diagram was and then name the part within the diagram. Moreover, the males may have an advantage because the diagram shows a male part.

Item 37, from the module 'Force', showed the largest DIF effect size: 448.91. The observed advantage of boys over girls on the module 'Force', concurred with other findings (Erickson & Erickson, 1984) that males out-performed females on items related to principles of gravitation and motion. The discrepancy may be explained

26. Which of the following parts bend light coming into the eye?

A. Iris
B. Lens
C. Retina
D. Pupil

51. Which of the following is not a safety device when using electricity?

A. Trip switch
B. Neutral wire
C. Fuse
D. Earth wire

FIG. 2. Content items.

by experiential differences. For instance, item 37 deals with the lifting of loads, which is more of a male task.

The last two items are characterised by content type. Items 51 and 26, from the module 'Energy', had significant DIF effect sizes of 85.22 and 114.65 respectively (see Fig. 2). Item 51 deals with electricity and the significant effect size concurs with other findings that males perform better than females in items related to electricity probably because the content area is drawn from their sphere of experience (Erickson & Erickson, 1984). Item 26 appears to be a straightforward recall of the functions of the parts of an eye and it is not clear why the boys did better than girls (see Fig. 2 for content items).

Girls' Performance. The six items are drawn from the modules 'Science in the home', 'Healthy living', 'Family life' and 'Energy'. Five of the Six items, 2, 28, 43, 49 and 56, flagged for DIF in favour of girls, are simple recall multiple-choice items of the 'correct answer' variety requiring ability to match and make associations.

Three of these Items, 2, 49 and 56, exhibit content that is useful for day-to-day living in Botswana, revealing that the perceived utility of the content plays a part in girls' performance. For instance, knowledge on first aid is useful for girls who, culturally, are regarded as home-makers, care-givers and nursemaids for younger siblings in the home. Girls, especially in the rural areas, help their mothers build mud huts and are more conversant than boys on basic needs in the home, such as energy. This finding concurs with other findings on gender differences in the goals for learning science. Charakupa (1995) found that the second ranked goal for learning science for females was improved safety awareness while the corresponding goal for males was desire for a science-based career. Gender differences in content performance would therefore also seem to depend on the perceived benefit of the knowledge gained (see Fig. 3 for utility of content items).

For items 28 and 43 the girls, in addition to using their ability to recall, were also able to use their ability to make associations to arrive at the correct answer (see Fig. 4 for association items). Item 8 (see Fig. 5), with the second largest DIF effect size, was context-dependent, drawing from the girls' sphere of interest. The girls did

2. Which of the following can be used to treat minor burns?

   A. First aid kit
   B. Fire extinguisher
   C. Safety goggles
   D. Clinical thermometer

49. Which of the following can be used as a building material?

   A. Coal
   B. Diamond
   C. Granite
   D. Soda ash

56. The most abundant and economic form of energy in Botswana is

   A. Wind energy
   B. Electrical energy
   C. Solar energy
   D. Tidal energy

FIG. 3. Utility of content items.

better in item 8 possibly because it involved their daily experiences. Girls would know from the ceremonies they have when they reach menarche that the first sign of pregnancy is missing a period. The phrases 'morning sickness' and 'period' are also clichés used and better understood within the female world of experiences with their body (see Fig. 5 for context-dependent items). It would seem that the largest DIF effect size is observed where ability to answer the item also depends on out-of-school experience.

Case-study 2: the Botswana Junior Certificate English Reading Passage, 1995

Method

The content of a Junior Certificate national examination English reading passage,

28. Which of the following pairs is correctly matched?

   A. Resistance — Watts
   B. Voltage — Ohms
   C. Current — Amperes
   D. Power — Volts

43. A salt produced from the reaction of nitric acid and magnesium is called

   A. Magnesium oxide
   B. Magnesium nitrite
   C. Magnesium nitride
   D. Magnesium nitrate

FIG. 4. Association items.

8. Lesego is 25 years old. She has missed her monthly period and has morning sickness. The most likely reason for this condition is that she

   A. Could be pregnant
   B. Could be reaching menopause
   C. Has sexually transmitted disease
   D. Has an irregular menstrual cycle

FIG. 5. Context-dependent item.

written in 1995, is analysed to demonstrate forms of gender bias that have implications for assessment. The candidates were required to read the passage and answer questions. Bias in the passage was reviewed by assessing gender representativeness, attributes assigned to the sexes, power relations and the use of language.

Illustrative Example of Gender-biased Language in an English Reading Passage (see Fig. 6)

Gender Representativeness. Gender representation was defined as the extent to which an item can be characterised as referring to or showing a male or female. For instance, pictorial items may be assessed to find out if males and females are equally shown in the pictures, while verbal items are investigated to find out if names and pronouns are used equally for all genders. The overall effect of the passage is biased because reference is predominantly made to men. The names of men and reference to men by their names or use of the pronoun 'he' or 'his' appear about 50 times compared with 15 references to women.

Attribution of Behaviours and Traits to One Sex. There is gender bias if certain behaviours, traits and attributes are portrayed as appropriate for one sex. In the passage, fearfulness and excitement are traits attributed to women while men are fearless. For instance, Ugoye is described as speechless with fear and now and again rubbing her hands together and holding them up to the sky. The rest of the women are described as following the men fearfully at a good distance and watching from afar, only to come running with excitement at the end.

Asymmetrical Dualism in the Relationship Between Women and Men. The dominant theme in the passage is the contrast in what the men and the women are doing. The men are more interested in the public sphere, where they are trying to untangle the tension between Christianity and the traditional beliefs. Ezeulu is glorified as a priest whose wisdom is unquestionable and unequivocal. The priest is an embodiment of culture and is almost a vanguard of that culture, not for his own sake but for the social good. Women are relegated to the domestic sphere of looking after the children and are equated with children, thus defining their standing in the social hierarchy. Even the boy children assume a higher social standing compared with their mothers.

'Father, Oduche's box is moving. It is moving about the floor,' Nwafo said out of breath with excitement.

'There is nothing a man will not hear nowadays!' exclaimed old Ezeulu in amazement but he stood very slowly to hide his curiosity. He went into the yard through the door at the back of his hut. Nwafo ran past him to the group of excited women outside his mother's hut. Akueke and Matefi did most of the talking. Nwafo's mother, Ugoye, was speechless with fear. Now and again she rubbed her hands together and held them up to the sky.

Akueke turned to Ezeulu as soon as she saw him 'Father, come and see. This new religion ...'

'Shut your mouth,' said Ezeulu who did not want anybody, least of all his daughter, to question his wisdom in sending one of his sons to join the new religion.

The wooden box had been brought from the room where Oduche and Nwafo slept and placed in the center of their mother's hut where people sat during the day.

The box, which was the only one of its kind in Ezeulu's compound, had a lock. Only people of the church had such boxes made for them by the mission carpenter and they were highly valued in Umuaro village. Oduche's box was not actually moving; but seemed to have something inside struggling to be free. Ezeulu stood before it wondering what to do. Whatever was inside the box became more and more violent and actually moved the box around. Ezeulu waited for it to calm down a little then he bent down, picked up the box and carried it outside. The women and children scattered in all directions.

'Now I shall see whether it contains good or bad medicine,' he said as he carried the box at arm's length like a sacrifice for the gods. His second son, Obika, followed him. Nwafo came closely behind Obika and the women and children followed fearfully at a good distance. Ezeulu looked back and asked Obika to bring him an axe. He took the box right outside his compound and finally put it down by the side of the footpath. 'Every one of you go back to the house,' he ordered. They moved back and stood in front of the hut. Obika took the axe to his father who thought for a while, put it aside and sent him for a long spear. The struggling in the box was as fierce as ever. For a brief moment Ezeulu wondered whether the wisest thing was not to leave the box there until its owner returned. But what would that mean? That he, Ezeulu, the priest of Umuaro was afraid of whatever his son had imprisoned within the box. Such a story must never be told.

He took the spear from Obika and pushed its thin end between the box and the lid. He clenched his teeth in an effort to open the box. The old priest was covered in sweat by the time he succeeded in forcing the box open. What they saw was enough to blind a man. Ezeulu stood speechless. The women who had watched from afar came running down. Soon a big crowd gathered. In the box, exhausted from its efforts, lay the priest's sacred python.

FIG. 6. Illustrative example of gender-biased language in an English reading passage (passage from *Arrow of God*, by Chinua Achebe).

Language. Gender bias in language may be revealed by the use of 'man' in the generic sense, for instance using 'he' 'him' 'his' as generic pronouns. It may also be revealed by the use of 'man' in idioms and some phrases, and 'man' as a suffix in occupational titles or specific groups of people, for example 'businessman'. It may also appear as a prefix in some words, such as 'mankind'. In the passage above (see Fig. 6), the word 'man' is used in a generic sense to refer to a person, for example 'there is nothing a man will not hear nowadays'. It is also used in an idiom, for example 'What they saw was enough to blind a man', even though the idiom is referenced to persons.

Implications of the Case-studies for Crafting Gender-fair Assessment

The case-study of the Junior Certificate science examination suggests that there may be numerous forms of knowledge patterned around male and female perceptions of reality. There appear to be three characteristics that account for the polarisation of item performance according to gender: content area and perceived benefit of the knowledge: item features; and out-of-school experience. The analysis showed that females were disadvantaged in two modules, 'How scientists work' and 'Force', while they performed better in the modules 'Science in the home', 'Family life', 'Healthy living' and 'Matter'. The items where girls' performance was superior to the boys, were drawn from modules that have a humane emphasis inclined towards nurturing and caring.

The question is whether such content areas should be eliminated from the examination paper. Eliminating content areas may threaten the curricular validity of the examination. DIF analysis, however, is essential to sensitise test development officers to the type of content that produces gender differences in performance. The role of the test development officer would therefore be to balance the content areas such that no group is disadvantaged. Nevertheless, further studies on DIF items in the science Junior Certificate examination need to be conducted before conclusions can be drawn on content areas that are biased against females.

The difference in item performance between boys and girls is also distinguished by item features. Three of the five items flagged for DIF in favour of boys were illustrated with diagrams, confirming the hypothesis that diagrammatic items introduce spatial ability competencies in which males out-perform females. The diagrams aid in assessing complex high-order thinking skills, thus avoiding unnecessary memorisation of specific content. The superior performance of males in these items would seem to suggest that males are better than females in high-order thinking skills. This may, however, be an unfair conclusion since most of these items used diagrams as background information. Interpretative materials described in prose need to be used to assess high-order skills so as to enable girls to take advantage of their verbal ability.

Five of the six items flagged for DIF in favour of girls are simple recall multiple-choice items of the 'correct answer' variety requiring ability to match and make associations. These items are also drawn from the modules 'Science in the home', 'Healthy living', 'Family life' and 'Energy'. Three of these items exhibit content that is useful for day-to-day living in Botswana, revealing that the perceived utility of the content plays a part in girls' performance. Gender differences in content performance would therefore also seem to depend on the perceived benefit of the knowledge gained. Item 8, with the second largest DIF effect, was, however, context-dependent, drawing from the girls' sphere of interest. It would seem that the largest DIF effect size is observed where ability to answer the item also depends on out-of-school experience.

Should examinations therefore avoid context-dependent items? Decontextualising assessment reduces the amount of assessment material available and may make assessment of higher-order thinking skills difficult (Gipps & Murphy, 1994). It also

may make generalisability of the examination to the out-of-school experience difficult, thus posing a threat to the construct validity of the examination. It is suggested that rather than decontextualise the examination, the tests should exhibit a variety of context that is reflective of the two groups. It is also proposed that where there seems to be a performance difference in content areas, the context of the item should be favourable to the disadvantaged group. Content areas that are too difficult for one group should be referred to curriculum development officers and test development officers for the development of enrichment and remedial materials for the disadvantaged group.

The second case-study, on the reading passage of the Junior Certificate English examination, has shown that examination materials are not necessarily gender-neutral. The language and the content communicate images, expectations and values that are gendered. The passage revealed bias in gender representativeness, attributes assigned to the sexes, power relations and the use of language. Gendered examination materials have implications for girls' and boys' performance as well as their socialisation. Gender-biased language may delay the processing of assessment tasks, resulting in low performance for the group affected. Examination materials also convey messages that may ultimately influence the way boys and girls perceive themselves. The reading passage, for instance, is demeaning to women and may be offensive to gender-sensitive examinees. Offensive materials may cause delays in the processing of the task, ultimately affecting performance in the whole test. Creating gender-free assessment materials amounts to decontextualising assessment. A strategy to achieve fairness may be to represent gender in a balanced and inoffensive way in the assessment materials (Nitko, 1996). Examination items on gender-biased materials could also require interpretations of the materials that reveal the prejudices so that assessment does not perpetuate gender stereotypes.

Conclusion

The article has assumed a bipolar duality in the practices of males and females that subsequently result in a gendered dichotomy of what counts as knowledge. It must, however, be mentioned that an emphasis on group differences may reinforce group stereotypes that may be harmful to the groups. The perspective also has the disadvantage of treating gender as a static concept. Yet sometimes this emphasis may be necessary in order to understand, recognise and make visible the multiple positions from which groups may perceive reality. In the case of assessment, the differences are necessary in order to recognise the hitherto undervalued females' ways of knowing so that these ways may be factored into our definitions of achievement. Such an approach requires that the meaning of achievement should be continuously revised so as to include the ever-changing perceptions of reality that come out throughout the life-long construction of gender.

There is thus a need for gender-fair assessment policies. Such policies should respond to research on gender differences in achievement so that definitions of achievement and the 'measuring stick of worth' can consistently be revised to ensure multiplicity of evidence on what has been learnt, and to provide enrichment

Subjects assessed

Which subjects offered in the curriculum count towards achievement?
What definitions of achievement inform the choices of subjects that count?
What body of knowledge concerning male and female differences informs the choices of these subjects?
Is there a gender balance in the subjects constituting achievement?

Assessment Task Formats

Are the assessment task formats exhaustive of the sources of evidence for the body of knowledge and skills reflected in the curriculum?
What body of knowledge concerning males' and females' differences informs evidence of task formats?
Using that body of knowledge, are the assessment task formats gender-balanced?

Content

What content areas are sampled for assessment and how do they relate to the evidence we have about gender differences in performance relative to content?
Are the content areas a representative sample of the curriculum domain?

Context-dependent items

To what extent can the interpretative element in assessment material be characterised as representing male or female experiences?
Are the forms of interpretative materials representative of male and female ways of knowing?
Is the situation in the interpretative material described in prose, presented in a graph or table or simulated by a picture, or does it require quantification?
On the basis of evidence of how context interacts with gender differences, curriculum experiences and out-of-school experiences, are the contexts chosen justifiable?
Is there a gender balance in the contexts selected?

Learning Outcomes

What weights are given to process and/or product and a combination as learning outcomes?
Has evidence on gender differences informed the decision on the weights?

Types of Ability

What taxonomies are used to categorise abilities?
Are the behaviourists as well as qualitative approaches to structuring knowledge used?

Gender Roles

Pictorial Items and Assessment Materials

Are males and females equally shown in the pictures?
Is the use of 'man' in the generic sense avoided?

Verbal Items and Assessment Materials

Are names and pronouns used equally for both sexes?
Are there any gender stereotypes portrayed by names?
Do the socio-cultural labels assigned to males and females give them an equal advantage?

Socio-cultural Scenario and Power Relations

Does the scenario represent a role stereotype?
Who is active? Who is passive or helpless?
Does the scenario equally enhance males and females? Do males and females equally exercise power and control in the scenario?

FIG. 7. Checklist for gender-fair assessment.

programmes in curriculum areas where one gender is disadvantaged. Policies on assessment should insist on aggregation of student performance in subjects and within subjects by test item by gender. It is only when such aggregation is mandatory that assessment can quickly respond to gender differences.

A number of examination boards have come up with guidelines for reviewing gender bias in national examinations. For example, the University of London School Examinations Board (1985) has produced guidelines on eliminating sexism and gender bias. Sexism, according to the guidelines, is shown through:

(1) maintenance of out-moded sexual stereotypes;
(2) use of language and illustrations;
(3) neglect of the contribution made by females to some subjects;
(4) reference to the interests of one sex rather than another (University of London School Examinations Board, 1985, p. 11).

The guidelines contend that gender bias is a statistical concept and that biased items may not be removed from the examination without threatening the construct validity of the curriculum. The suggestion made is that there must be an attempt to produce a balance in the overall assessment.

The article has revealed that gender interacts with numerous forms of knowledge. For instance, gender interacts with subject choice. However, within a subject, numerous forms of knowledge emerge that also interact with gender, for instance task formats used, content areas sampled, context of the assessment task and types of abilities assessed. The forms of knowledge discussed are not exhaustive, an indication that there is still a lot more research needed in order to understand gendered effects on achievement. Frameworks to assist in producing a gender balance in assessment based on existing evidence on gender and achievement are nevertheless necessary to achieve fairness. Figure 7 gives a checklist that may assist in crafting gender-fair assessment.

References

ANGOFF, W. F. (1982) Use of difficulty and discrimination indices for detecting item bias, in: R. A. BERK (Ed.) *Handbook Methods for Detecting Item Bias*, pp. 96–116 (Baltimore, MD, Johns Hopkins University).

BATESON, D. J. & PARSONS-CHATMAN, S. (1989) Sex related differences in science achievement: a possible testing artefact, *International Journal of Science Education*, 11, pp. 371–385.

BEN-SHAKAR, G. & SINAI, Y. (1991) Gender and guessing, *Journal of Educational Measurement*, 28, pp. 23–35.

BLOOM, B. S., ENGELHART, M. D., FURST, E. J., HILL, W. H. & KRATHWOHL, D. R. (1956) *Taxonomy of Educational Objectives: the classification of educational goals. Handbook 1: Cognitive domain* (New York, Longman).

CHARAKUPA, R. (1995) Learning science: revisiting the goals and gender issues at the community junior secondary schools in Botswana, *Southern African Journal of Science and Mathematics Education*, 2, pp. 101–121.

CHILISA, B. M. (1997) Ecological and gender bias in science achievement: can judgmental and statistical procedures for eliminating bias produce fair tests? A paper presented at the 23rd annual International Association for Educational Assessment conference, Durban, South Africa.

CHILISA, B. M. & MATAMBO, N. (1997) *Investigating Gender Bias in the Junior National Examinations. A Research Report* (Gaborone University of Botswana).

CRAWFORD, M. & ENGLISH, L. (1984) Generic versus specific inclusion of women in language: effects on recall, *Journal of Psycholinguistic Research*, 13, pp. 373–381.

DORANS, N. J. (1989) Two new approaches to assessing differential item functioning: standardisation and the Mantel–Haenszel method, *Applied Measurement in Education*, 2, pp. 217–233.

EICHLER, M. (1991) *Non-sexist Research Methods: a practical guide* (New York, Routledge).

ERICKSON, J. L. & ERICKSON, L. J. (1984) Females and science achievement: evidence, explanations and implications, *Science Education*, 68, pp. 63–89.

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

HAMBLETON, R. K. & ROGERS, H. J. (1989) Detecting potentially biased items: comparison of IRT and Mantel–Haenszel methods, *Applied Measurement in Education*, 2, pp. 313–334.

HARDING, S. (1986) *The Science Question in Feminism* (New York, Cornell University Press).

HENLEY, M. (1989) Molehill or mountain? What we know and don't know about sex bias in language, in: M. CRAWFORD & M. GENTRY (Eds) *Gender and Thought: psychological perspectives*, pp. 58–78 (New York, Springer-Verlag).

HILDEBRAND, G. M. (1996) Redefining achievement, in: P. F. MURPHY & C. GIPPS (Eds) *Equity in the Classroom: towards effective pedagogy for girls and boys*, pp. 149–172 (London, Falmer Press).

HOLLAND, P. W. & THAYER, D. T. (1988) Differential item functioning and the Mantel–Haenszel procedure, in: H. WAINER & H. BRAUN (Eds) *Test Validity*, pp. 129–172 (New York, Hillsdale).

KIMURA, D. (1992) Sex differences in the brain, *Scientific America*, 267, pp. 81–87.

LEVI-STRAUSS, C. (1969) *The Elementary Structures of Kinship* (Boston, MA, Beacon Press).

LEVY, J. (1980) Cerebral asymmetry and the psychology of man, in: M. C. WITTROCK (Ed.) *Brain and Psychology*, pp. 51–651 (Orlando, FL, Academic Press).

MACCOBY, E. E. & JACKLIN, C. N. (1974) *The Psychology of Sex Differences* (Stanford, CA, Stanford University Press).

MOORAD, F. R. (1993) Development of education in Botswana: focus on community initiatives, *Mosenodi Journal of the Botswana Educational Research*, 1, pp. 31–47.

NITKO, A. J. (1996) *Educational Assessment of Students* (Englewood Cliffs, New Jersey, Prentice-Hall).

ORTNER, S. B. (1972) Is female to male as nature is to culture? Feminist studies, in: M. Z. ROSALDO & L. LAMPHERE (Eds) *Woman, Culture and Society*, pp. 67–88 (Stanford, CA, Stanford University Press).

ORTNER, S. B. & WHITEHEAD, H. (1991) Accounting for sexual meanings, in: S. B. ORTNER & H. WHITEHEAD (Eds) *Sexual Meanings: the cultural construction of gender and sexuality*, pp. 1–261 (Cambridge, Cambridge University Press).

ORTON, A. (1987) *Learning Mathematics: issues, theory and classroom practice* (London, Cassell).

PERRY, W. (1970) *Forms of Intdlectual and Ethical Development in the Colleges* (Orlando, FL, Harcourt Brace College Publishers).

RENNIE, L. J. & PARKER, L. H. (1991) Assessment of learning in science: the need to look closely at item characteristics, *Australian Science Teachers' Journal*, 37(4), pp. 56–59.

ROSALDO, M. Z. (1974) Woman, culture and society: a theoretical overview, in: M. Z. ROSALDO & L. LAMPHERE (Ed.) *Womart, Culture and Society*, pp. 17–42 (Stanford, CA, Stanford University Press).

SHAPERA, I. (1959) *Tswana Law and Custom* (London, Frank Cass).

SHEMESH, M. (1990) Gender-related differences in reasoning skills and learning interests of junior high school students, *Journal of Research in Science Teaching*, 27, pp. 27–34.

SHEPARD, L. A., CAMILLI, G. & WILLIAMS, D. M. (1985) Validity of approximation techniques for defecting item bias, *Journal of Educational Measurement*, 22, pp. 77–105.

SILVEIRA, J. (1978) Women on the fringes: generic masculine words and their relationship to thinking, unpublished manuscript.

STOBART, G., ELWOOD, J. & QUINLAN, M. (1992) Gender bias in examinations: how equal are the opportunities? *British Educational Research Journal*, 18, pp. 261–276.

TAIWO, A. A. & MOLOBE, E. N. (1994) Gender dimensions of the perceptions of subject and career choices of students: a case study of Botswana senior secondary schools, *Southern Africa Journal of Mathematics and Science Education*, 1, pp. 3–23.

THOMPSON, A. & GITLIN, A. (1995) Creating space for reconstructing knowledge in feminist pedagogy, *Educational Theory*, 45, pp. 145–165.

TOBIAS, E. S. (1990) *They Are Not Dumb, They Are Different* (Tucson, Research Corporation).

TOWNSEND-COLES, E. K. (1985) *Education in Botswana* (Gaborone, Macmillan).

UNIVERSITY OF LONDON SCHOOL EXAMINATIONS BOARD (1985) *Sexism, Discrimination and Gender Biases in GCE Examinations* (London, University of London).

ZWICK, R. & ERICKAN, K. (1989) Analysis of the differential item functioning in the NAEP history assessment, *Journal of Educational Measurement*, 26, pp. 55–66.