2 Test fairness

Antony John Kunnan
California State University, USA

Abstract

The concept of test fairness is arguably the most critical in test evaluation but there is no coherent framework that can be used for evaluating tests and testing practice. In this paper, I present a Test Fairness framework that consists of the following test qualities: validity, absence of bias, access, administration, and social consequences. Prior to presenting this framework, I discuss early views of test fairness, test evaluation in practice and ethics for language testing. I conclude with some practical guidelines on how the framework could be implemented and a discussion of the implications of the framework for test development.

Introduction

The idea of test fairness as a concept that can be used in test evaluation has become a primary concern to language-testing professionals today, but it may be a somewhat recent preoccupation in the history of testing itself. Perhaps this is so because of the egalitarian view that tests and examinations were considered beneficial to society, as they helped ensure equal opportunity for education and employment and attacked the prior system of privilege and patronage. For this reason, tests and examinations have taken on the characteristic of infallibility. But everyone who has taken a test knows that tests are not perfect; tests and testing practices need to be evaluated too.

The first explicit, documented mention of a test quality was in the 19th century, after competitive examinations had become entrenched in the UK. According to Spolsky (1995), in 1858 a committee for the Oxford examinations 'worked with the examiners to ensure the general *consistency* of the examination as a whole' (p. 20). According to Stigler (1986), Edgeworth articulated the notion of *consistency* (*or reliability*) in his papers on error and chance much later, influenced by Galton's anthropometric laboratory for studying physical characteristics. As testing became more popular in the later decades of the 19th century and early 20th century, modern measurement

theory (with influential treatments from Galton, Edgeworth, Spearman, Pearson, Stevens, Guilford and Thurstone) developed techniques including correlation and factor analysis. These statistical procedures became the primary evaluative procedures for test development and test evaluation. In modern language assessment, test evaluation has clearly derived from this tradition of statistical procedures (and quantitative methods). In recent years there has been interest in using qualitative methods, and the concept of fairness too has specifically emerged, but a framework that includes these methods and concepts has not been debated.

In this paper, influenced by the work of Messick on test validation, and by ethics and philosophy, I present a framework of test fairness that broadens the scope of traditional test evaluation.

Early approaches to test evaluation

Many testing professionals hold the view that testing research has always focused on issues of fairness (and related matters such as bias, justice and equality), within the framework of test evaluation through the concepts of validity and reliability. A closer examination of this view should clarify whether this is an acceptable idea. Influenced by the work in statistics and measurement and the *Standards* (actually recommendations) for educational and psychological tests and manuals of the American Psychological Association (1954), Lado (1961) was the first author in modern language assessment to write about test evaluation in terms of validity (which encompasses face validity, validity by content, validation of the conditions required to answer the test items, and empirical validation in terms of concurrent and criterion-based validation) and reliability. Later, Davies (1968) presented a scheme for determining validities which listed five types of validity – face, content, construct, predictive and concurrent – and Harris (1969) urged test writers to establish the characteristics of a good test by examining tests in terms of content, empirical (predictive and concurrent), and face validity.

The *Standards* were reworked during this time (APA 1966, 1974) and the interrelatedness of the three different aspects of validity (content, criterion-related and construct validities) was recognised in the 1974 version. This trinitarian doctrine of content, criterion-related and construct validity (reduced in number because the concurrent and predictive validity of the 1954 version were combined and referred to as criterion-related) continued to dominate the field. In 1985, the *Standards* were reworked again and titled 'Standards for educational and psychological test' (instead of *Standards for tests*).[1] This new reworking included Messick's unified and expanded conceptual framework of validity, which was fully articulated in Messick (1989) with attention to values and social consequences of tests and testing as facets of validity of test-score interpretation. But most books in language testing did not present Messick's

unified and expanded view of validity (see Henning 1987; Hughes 1989; Alderson, Clapham and Wall 1995; Genesee and Upshur 1996; and Brown 1996). Only Bachman (1990) presented and discussed Messick's unified and expanded view of validity. Thus, validity and reliability continue to remain the dominant concepts in test evaluation, and fairness has remained outside the mainstream.

The 'Test Usefulness' approach to test evaluation

The 1990s brought a new approach to test evaluation. Translating Messick's conceptual framework, Bachman and Palmer (1996, B and P hereafter) articulated their ideas regarding test evaluation qualities: 'the most important consideration in designing and developing a language test is in the use for which it is intended, so that the most important quality of a test is its usefulness' (p. 17). They expressed their notion thus: 'Usefulness = Reliability + Construct Validity + Authenticity + Interactiveness + Impact + Practicality' (p. 18). This representation of test usefulness, they asserted, 'can be described as a function of several different qualities, all of which contribute in unique but interrelated ways to the overall usefulness of a given test' (p.18). The B and P approach does not directly address the concept of fairness but they do show an awareness of it in their use of Messick's expanded view of validity and their inclusion of impact as one of the 'test usefulness' qualities.

Test evaluation in practice

Another way of noting which test evaluation qualities were important to researchers is to examine the research they carried out. For example, researchers at Educational Testing Service, Princeton, examined the TOEFL, the TSE and the TWE, and produced 78 research and technical reports (55 on the TOEFL, 11 on TSE, and 12 on TWE). The areas of inquiry include test validation, test information, examinee performance, test use, test construction, test implementation, test reliability, and applied technology (ETS 1997).[2] The University of Cambridge Local Examinations Syndicate, UK (UCLES), which administers many EFL tests including the FCE, CPE and the IELTS, examined their tests too but judging from the IELTS research reports (IELTS 1999), the range of studies is limited to investigations of test reliability, validity and authenticity.[3,4] The English Language Institute, University of Michigan, Ann Arbor, which administers many EFL tests, produced a technical manual in support of the Michigan English Language Assessment Battery.[5] This manual includes discussions on validity and reliability using quantitative methods.[6]

The 13th edition of the *Mental Measurement Yearbook* (Impara and Blake 1998; MMY for short) has 693 reviews of 369 tests and includes reviews of 21 tests in English, 13 in reading, and four in foreign languages. Of these 38

test reviews, most of them uniformly discuss the five kinds of validity and reliability (typically, in terms of test-retest and internal consistency), and a few reviews discuss differential item functioning and bias.[7] The 10th Volume of *Test Critiques* (Keyser and Sweetland 1994) has 106 reviews which include seven related to language. Although the reviews are longer and not as constrained as the ones in the MMY, most reviews only discuss the five kinds of validity and the two kinds of reliability. The *Reviews of English Language Proficiency Tests* (Alderson *et al.* 1987) is the only compilation of reviews of English language proficiency tests available. There are 47 reviews in all and they follow the MMY's set pattern of only discussing reliability and validity, mostly using the trinitarian approach to validity while a few reviews also include discussions of practicality. There is no direct reference to test fairness.

Early test bias studies

While these general studies and reviews do not as a rule focus on the concept of fairness, a separate interest in developing culture and bias-free tests developed in educational testing. These studies began with the narrow focus on test and item-bias studies and then developed into the technical literature now known as DIF (Differential Item Functioning) studies.[8] Early landmark studies, cited in Willingham and Cole (1997), examined predictions of grades for black and white college students (Cleary 1968, and Cleary and Hilton 1969), differential validity of employment tests by race (Hunter, Schmidt and Hunter 1979), and fair selection models (Cole 1973; Cole and Moss 1989). Similarly, in language-testing research in the last two decades, gender, academic major, and native language and culture group differences have been examined the most (examples: Chen and Henning 1985; Alderson and Urquhart 1985ab; Zeidner 1986, 1987; Oltman *et al.* 1988; Hale 1988; Kunnan 1990; Ryan and Bachman 1992; Bachman *et al.* 1995; Kunnan 1995; Elder 1996; Clapham 1996, 1998; Ginther and Stevens 1998). Other relevant studies include examination of washback (Wall and Alderson 1993), test-taker feedback (Norton and Stein 1998), and test access (Brown 1993; Taylor *et al.* 1998). In summary, while some researchers are interested in test bias, the approach is fragmentary at best and not all tests are evaluated using a fairness framework.

In conclusion: first, although this overall 'engineering' approach, greatly influenced by the invention of statistical techniques, helped provide the tools necessary for validity and reliability studies, this unfortunately made most researchers complacent as they valued only statistical evidence and discounted other types of investigations and evidence. Second, single narrow-scope studies give the illusion that they have accomplished more than they set out to do. For example, a single DIF study (investigating gender differences in performances) might typically attempt to provide answers to the question of

test or item bias for certain groups, but might not be able to answer questions regarding other group differences. Or a single validation study (of, say, internal structure), while useful in its own right, would have insufficient validation evidence to claim that the test has all the desirable qualities. Third, published test reviews are narrow and constrained in such a way that none of the reviews I surveyed follow Messick's (1989) concepts of test interpretation and use and evidential and consequences bases of validation, and, therefore, they do not provide any guidance regarding these matters. In short, based on the analyses above, test evaluation is conducted narrowly and focuses mainly on validity and reliability.

Ethics in language testing

A language-test ethic has been slow to develop over the last 100 years. Spolsky (1995) convincingly argued that from the 1910s to the 1960s, social, economic and political concerns among key language-testing professions in the US (mainly at the Ford Foundation, the College Board, and Educational Test Service, Princeton) and the UK (mainly at the University of Cambridge Local Examinations Syndicate [UCLES]) dominated boardroom meetings and decisions. A language-test ethic was not evident in this period although ethical theories of different persuasions had been in existence for several centuries.

Davies (1977) was the first to make an interesting suggestion for 'test virtues' which could be seen as the first suggestion of ethical concerns in language testing. Except for Davies' 'test virtues' of reliability and validity, there has been no mention of test ethics. In the last two decades, ethical concerns emerged sporadically in language assessment. Spolsky (1981) argued that tests should be labelled, like drugs, 'Use with care'. Stevenson (1981) urged language testers to adhere to test development standards that are internationally accepted for all educational and psychological measures. Canale (1988) suggested a naturalistic-ethical approach to language testing, emphasising that language testers should be responsible for ethical use of the information they collect. Groot (1990) argued for checks on the quality of tests as essential for valid conclusions and decisions. Stansfield (1993) argued that professional standards and a code of practice are ways of bringing about ethical behaviour among testers. Alderson *et al.* (1995) reviewed principles and standards but concluded that 'language testing still lacks any agreed standards by which language tests can be evaluated, compared or selected' (p. 259).

Broadening the discussion, Cumming (1994) reviewed the functions of language assessment for recent immigrants to Canada and asked a fundamental question: 'Does the process of language assessment help or hinder...?' (p. 117). He raised three problems regarding the way in which language assessment functions within Canadian society: language assessment

may pose barriers to recent immigrants' participation in Canadian society; it may be too limited in scope; and it may put the burden of responsibility onto the performance of individual immigrants. Valdés and Figueroa (1994) addressed the reasons underlying bilingual children's poor performance on standardised tests, arguing that without an understanding of the nature of bilingualism itself, the problems encountered by bilingual individuals on such tests will continue. The last two studies began to consider language assessment from a broader, societal perspective, which features ethical issues such as justice, equity and participation.

In the last few years, momentum has gathered through publications such as the special issue of *Language Testing* guest-edited by Davies (1997a), which contained significant papers by Spolsky (1997), Lynch (1997), Hamp-Lyons (1997a), Shohamy (1997) and Davies (1997b), and other important papers by Hamp-Lyons (1997b) and Norton (1997). The International Language Testing Association (ILTA) recently published a report by the Task Force on Testing Standards (1995), which was followed by ILTA's Code of Ethics (2000) which lays out some broad guidance on how professionals should conduct themselves. However, these documents are general explorations of applied ethics and lack specific application of ethical methods which can be applied to test evaluation.

The three predominant, secular, ethical methods (utilitarianism, Kantian and deontological systems, and virtue-based ethics) may give us some guidance. To elaborate briefly, *utilitarianism*, which emphasises good results as the basis for evaluating human actions, has two main features: its teleological aspect or *consequentialist principle* and the hedonic aspect or *utility principle*. In the words of Pojman (1999), 'the consequentialist principle states that the rightness or wrongness of an act is determined by the goodness or badness of the results that flow from it. It is the end, not the means, that counts; the end justifies the means. The utility principle states that the only thing that is good in itself is some specific types of state (e.g. pleasure, happiness, welfare)' (p. 109). [9 10]

In contrast, *Kantian* or *deontological* ethics focuses on ideals of universal law and respect for others as a basis of morality and sees the rightness or wrongness of an act in itself, not merely in the consequences of the act. In other words, the end never justifies the means. But the two kinds of deontological theory vary slightly. Act-deontologists (who are intuitionists) believe that we must consult our conscience regarding every act in order to discover whether that act is morally right or wrong. Rule-deontologists (like Kant), on the other hand, accept the notion of universal principles and believe that when we make moral decisions we are appealing to rules. [11]

Virtue-based ethics, which views moral questions from the standpoint of the moral agent with virtuous characters, has re-emerged owing to dissatisfaction that may have arisen with the previous two methods. But

virtue-based ethics calls persons to be virtuous by possessing both moral and non-moral virtues by imitation, even though there are no principles to serve as criteria for the virtues.[12]

Whatever the methodological persuasion, language-testing professionals need an ethic to support a framework of applied ethical principles that could guide professional practice. Keeping these methods in mind, we could begin to consider Hamp-Lyons' (1997b) question (with slight modification): 'What is the principle against which the ethicality of a test is to be judged?' (p. 326). Corson (1997), broadly addressing applied linguists, makes a case for the development of a framework of ethical principles by considering three principles: the principle of equal treatment, the principle of respect for persons, and the principle of benefit maximisation.

In addition to Hamp-Lyons' question cited above, we need help with other sub- or auxiliary questions, such as: What qualities should a language test have in order to be considered an ethically fair test? What are the required qualities for a language-testing practice (meaning the rights and responsibilities of all stakeholders in a test, including the test developer, test user and test taker) in order for it to be considered one with fairness or right conduct? What qualities should a code of ethics and a code of practice have so that language assessment professionals can follow ethical practice?

An ethics-inspired rationale for the Test Fairness framework

I present an ethics-inspired rationale for my test fairness framework, with a set of principles and sub-principles. The principles are based on Frankena's (1973) 'mixed deontological' system, which combines both the utilitarian and the deontological systems. Frankena suggests reconciling the two types of theory by accepting the notice of rules and principles from the deontological system but rejecting its rigidity, and by using the consequential or teleological aspect of utilitarianism but without the idea of measurement of goodness, alleviation of pain, or to bring about the greatest balance of good over evil.

Thus two general principles of justice[13] and beneficence (plus sub-principles) are articulated as follows:

Principle 1: *The Principle of Justice*: A test ought to be fair to all test takers; that is, there is a presumption of treating every person with equal respect.[14]

Sub-principle 1: A test ought to have comparable construct validity in terms of its test-score interpretation for all test takers.

Sub-principle 2: A test ought not to be biased against any test-taker groups, in particular by assessing construct-irrelevant matters.

Principle 2: *The Principle of Beneficence*: A test ought to bring about good in society; that is, it should not be harmful or detrimental to society.

Sub-principle 1: A test ought to promote good in society by providing test-score information and social impacts that are beneficial to society.

Sub-principle 2: A test ought not to inflict harm by providing test-score information or social impacts that are inaccurate or misleading.

Test fairness framework

Defining Fairness

The notion of test fairness has developed in so many ways that the various positions may appear contradictory. One useful way of understanding the many points of view is to examine recent documents that have brought this to the forefront: the Code of Fair Testing Practices in Education (1988; *Code* for short) from the Joint Committees on Testing Practices in Washington, DC and the Standards (1999, *Standards* for short) for educational and psychological testing prepared by the American Educational Research Association, American Psychological Association and the National Council on Measurement in Education.

The *Code* approach

The *Code* (1988) presents standards for educational test developers and users in four areas: developing and selecting tests, interpreting scores, striving for fairness and informing test takers. Specifically, the *Code* provides practical guidelines for test developers and users on how to strive for fairness. Keeping these guidelines in mind, standards for implementation and acceptability for the qualities are discussed here. Here is the excerpt from Section C, *Striving for Fairness,* divided into two parts, one for test developers and one for test users:

> **Test developers** should strive to make tests that are as fair as possible for test takers of different races, gender, ethnic backgrounds, or handicapping conditions.
>
> **Test developers should:**
> Review and revise test questions and related materials to avoid potentially insensitive content or language.
> - Investigate the performance of test takers of different races, gender and ethnic backgrounds when samples of sufficient size are available. Enact procedures that help to ensure that differences in performance are related primarily to the skills under assessment rather than to irrelevant factors.
> - When feasible, make appropriately modified forms of tests or administration procedures available for test takers with handicapping conditions. Warn test users of potential problems in using standard norms with modified tests or administration procedures that result in non-comparable scores.

Test users should select tests that have been developed in ways that attempt to make them as fair as possible for test takers of different races, gender, ethnic backgrounds, or handicapping conditions.

Test users should:

- Evaluate the procedures used by test developers to avoid potentially insensitive content or language.
- Review the performance of test takers of different races, gender and ethnic backgrounds when samples of sufficient size are available. Evaluate the extent to which performance differences might have been caused by inappropriate characteristics of the test.
- When necessary and feasible, use appropriately modified forms of tests or administration procedures for test takers with handicapping conditions. Interpret standard norms with care in the light of the modifications that were made.

(Code 1988, p. 4–5)

The *Standards* (1999) approach

In the recent *Standards* (1999), in the chapter entitled, 'Fairness in testing and test use', the authors state by way of background that the 'concern for fairness in testing is pervasive, and the treatment accorded the topic here cannot do justice to the complex issues involved. 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' (p. 73). Furthermore, the document acknowledges the difficulty of defining fairness: 'the term *fairness* is used in many different ways and has no single meaning. It is possible that two individuals may endorse fairness in testing as a desirable social goal, yet reach quite different conclusions' (p. 74). With this caveat, the authors outline four principal ways in which the term is used[15]:

> The first two characterisations... relate fairness to *absence of bias* and to *equitable treatment of all examinees* in the testing process. There is broad consensus that tests should be free from bias... and that all examinees should be treated fairly in the testing process itself (e.g. afforded the same or comparable procedures in testing, test scoring, and use of scores). The third characterisation of test fairness addresses *the equality of testing outcomes* for examinee subgroups defined by race, ethnicity, gender, disability, or other characteristics. The idea that fairness requires equality in overall passing rates for different groups has been almost entirely repudiated in the professional testing literature. A more widely accepted view would hold that examinees of equal standing with respect to the construct the test is intended to measure should on average earn the same test score, irrespective of group membership... The fourth definition of fairness relates to *equity in opportunity to learn* the material covered in an

> achievement test. There would be general agreement that adequate opportunity to learn is clearly relevant to some uses and interpretations of achievement tests are clearly irrelevant to others, although disagreement might arise as to the relevance of opportunity to learn to test fairness in some specific situations.
>
> (*Standards* 1999, p. 74; emphasis added)

In addition, the document discusses two other main points: bias associated with test content and response processes, and fairness in selection and prediction. Based on these discussions, the document goes on to formulate twelve standards for fairness. The relevant standards are summarised here:

- Validity evidence collected for the whole test group should also be collected for relevant sub-groups.
- A test should be used only for the sub-groups for which evidence indicates that valid inferences can be drawn from test scores.
- When DIF exists across test-taker characteristic groups, test developers should conduct appropriate studies.
- Test developers should strive to identify and eliminate language and content that are offensive by sub-groups except when necessary for adequate representation of the domain.
- When differential prediction of a criterion for members of different sub-groups are conducted, regression equations (or appropriate equivalent) should be computed separately for each group.
- When test results are from high-stakes testing, evidence from mean score differences between relevant sub-groups should be examined and if such differences are found, an investigation should be undertaken to determine that such differences are not attributable to a source of construct under-representation or construct-irrelevance variance.

Willingham and Cole (1997) approach

Independent researchers like Willingham and Cole (1997) (in their study of gender and fair assessment) and Willingham (1999), emphasised several varying ideas in describing a system for considering fairness issues. They state that 'test fairness is an important aspect of validity... anything that reduces fairness also reduces validity... test fairness is best conceived as comparability in assessment; more specifically, comparable validity for all individuals and groups' (pp. 6–7). Using the notion of comparable validity as the central principle, Willingham suggests three criteria for evaluating the fairness of a test: 'comparability of opportunity for examinees to demonstrate relevant proficiency, comparable assessment exercises (tasks) and scores, and comparable treatment of examinees in test interpretation and use' (p. 11).

Based on these ideas, four characteristics of fairness emerge that are the most critical to fair assessment practices. They are: comparable or equitable treatment in the testing process, comparability or equality in outcomes of

learning and opportunity to learn, absence of bias in test content, language and response patterns, and comparability in selection. It is these characteristics that form the backbone of the framework that I propose below.

The Test Fairness framework

The Test Fairness framework views fairness in terms of the whole system of a testing practice, not just the test itself. Therefore, following Willingham and Cole (1997), multiple facets of fairness that includes multiple test uses (for intended and unintended purposes), multiple stakeholders in the testing process (test takers, test users, teachers and employers), and multiple steps in the test development process (test design, development, administration and use) are implicated. Thus, the model has five main qualities: validity, absence of bias, access, administration, and social consequences. Table 1 (see page 46) presents the model with the main qualities and the main focus for each of them. A brief explanation of the qualities follows:

1 **Validity**: Validity of a test score interpretation can be used as part of the test fairness framework when the following four types of evidence are collected.
   a) *Content representativeness or coverage evidence*: This type of evidence (sometimes simply described as *content validity*) refers to the adequacy with which the test items, tasks, topics and language dialect represent the test domain.
   b) *Construct or theory-based validity evidence*: This type of evidence (sometimes described as *construct validity*) refers to the adequacy with which the test items, tasks, topics and language dialect represent the construct or theory or underlying trait that is measured in a test.
   c) *Criterion-related validity evidence*: This type of evidence (sometimes described as *criterion validity*) refers to whether the test scores under consideration meet criterion variables such as school or college grades and on the job-ratings, or some other relevant variable.
   d) *Reliability*: This type of evidence refers to the reliability or consistency of test scores in terms of consistency of scores on different testing occasions (described as *stability* evidence), between two or more different forms of a test (*alternate form* evidence), between two or more raters (*inter-rater* evidence), and in the way test items measuring a construct functions (*internal consistency* evidence).

2 **Absence of bias**: Absence of bias in a test can be used as part of the test fairness framework when evidence regarding the following is collected.
   a) *Offensive content or language*: This type of bias refers to content that is offensive to test takers from different backgrounds, such as stereotypes of group members and overt or implied slurs or insults (based on gender,

race and ethnicity, religion, age, native language, national origin and sexual orientation).

b) *Unfair penalisation based on test taker's background*: This type of bias refers to content that may cause unfair penalisation because of a test taker's group membership (such as that based on gender, race and ethnicity, religion, age, native language, national origin and sexual orientation).

c) *Disparate impact and standard setting*: This type of bias refers to differing performances and resulting outcomes by test takers from different group memberships. Such group differences (as defined by salient test-taker characteristics such as gender, race and ethnicity, religion, age, native language, national origin and sexual orientation) on test tasks and sub-tests should be examined for Differential Item/Test Functioning (DIF/DTF)[16]. In addition, a differential validity analysis should be conducted in order to examine whether a test predicts success better for one group than for another. In terms of standard-setting, test scores should be examined in terms of the criterion measure and selection decisions. Test developers and users need to be confident that the appropriate measure and statistically sound and unbiased selection models are in use[17]. These analyses should indicate to test developers and test users that group differences are related to the abilities that are being assessed and not to construct-irrelevant factors.

3 **Access**: Access to a test can be used as part of the test fairness framework when evidence regarding the following provisions is collected.

a) *Educational access*: This refers to whether or not a test is accessible to test takers in terms of *opportunity to learn* the content and to become familiar with the types of task and cognitive demands.

b) *Financial access*: This refers to whether a test is *affordable* for test takers.

c) *Geographical access*: This refers to whether a test site is accessible in terms of distance to test takers.

d) *Personal access* here refers to whether a test provides certified test takers who have physical and/or learning disabilities with appropriate test accommodations. The 1999 *Standards* and the *Code* (1988) call for accommodation to be such that test takers with special needs are not denied access to tests that can be offered without compromising the construct being measured.

e) *Conditions or equipment access*: This refers to whether test takers are familiar with the test taking equipment (such as computers), procedures (such as reading a map), and conditions (such as using planning time).

4 **Administration**: Administration of a test can be used as part of the test fairness framework when evidence regarding the following conditions is collected:

a) *Physical conditions*: This refers to appropriate conditions for test administration, such as optimum light and temperature levels and facilities considered relevant for administering tests.
b) *Uniformity or consistency*: This refers to uniformity in test administration exactly as required so that there is uniformity and consistency across test sites and in equivalent forms, and that test manuals or instructions specify such requirements. Uniformity refers to length, materials and any other conditions (for example, planning time or the absence of planning time for oral and written responses) so that test takers (except those receiving accommodations due to disability) receive the test under the same conditions. Test security is also relevant to this quality, as a test's uniformity is contingent upon it being administered in secure conditions.

5 **Social consequences**: The social consequences of a test can be used as part of the test fairness framework when evidence regarding the following is collected:

a) *Washback*: This refers to the effect of a test on instructional practices, such as teaching, materials, learning, test-taking strategies, etc.
b) *Remedies*: This refers to remedies offered to test takers to reverse the detrimental consequences of a test, such as re-scoring and re-evaluation of test responses, and legal remedies for high-stakes tests. The key fairness questions here are whether the social consequences of a test and/or the testing practices are able to contribute to societal equity or not and whether there are any pernicious effects due to a particular test or testing programme [18].

In summary, these five test fairness qualities (validity, absence of bias, access, administration and social consequences), when working together, could contribute to fair tests and testing practices. Furthermore, the test fairness framework meets the guidelines of fairness in assessment contained in the recent *Code* (1988) and the *Standards* (1999). Finally, it is expected that the framework will be used in a unified manner so that a fairness argument such as the validity argument proposed by Kane (1992) can be used in defending tests as fair.

Implications for test development

The implications of the model for test development are significant. This is because the concern for fairness in language testing cannot be raised only after the test is developed and the test administered. The concern has to be present at all stages of test development: design, development, piloting and administration, and use (which includes analysis and research), although different fairness qualities may be in focus at different stages (Kunnan 2000).

Another way forward in test development is for the participants to recruit

test developers (thinkers or planners, writers, raters and researchers) from diverse groups (in terms of gender, race/ethnicity, native language, age, etc.) for training in fairness issues prior to test development. This could ensure that the many aspects of fairness would be well understood by the test developers.

Finally, the question of who is responsible for fairness-testing practices is worth raising: should it be the test developer or the test user, or both? In my view, as the two primary stakeholders of every test, both groups of individuals should be held responsible for promoting fairness.

Conclusion

In conclusion, this paper argues for a test fairness framework in language testing. This conceptualisation gives primacy to fairness and, in my view, if a test is not fair there is little value in a test having qualities such as validity and reliability of test scores. Therefore, this model consists of five interrelated test qualities: validity, absence of bias, access, administration, and social consequences.

The notion of fairness advanced here is based on the work of the *Code* (1988), the *Standards* (1999), and Willingham and Cole's (1997) notion of 'comparable validity'. This framework also brings to the forefront two qualities (*access* and *administration*) that are ignored or suppressed in earlier frameworks, as these qualities have not been seen as part of the responsibility of test developers. They have generally been delegated to test administrators and local test managers, but I propose that these two qualities should be monitored in the developmental stages and not left to the test administrators.

This framework, then, is a response to current concerns about fairness in testing and to recent discussions of applied ethics relevant to the field. Its applicability in varied contexts for different tests and testing practices in many countries would be a necessary test of its robustness. Further, I hope that the framework can influence the development of shared operating principles among language assessment professionals, so that fairness is considered vital to the professional and that societies benefit from tests and testing practices.

To sum up, as Rawls (1971) asserted, one of the principles of fairness is that institutions or practices must be *just*. Echoing Rawls, then, there is no other way to develop tests and testing practice than to make them such that primarily there is fairness and justice for all. This is especially true in an age of increasingly information-technology-based assessment, where the challenge, in Barbour's (1993) words, would be to 'imagine technology used in the service of a more just, participatory, and sustainable society on planet earth' (p. 267).

References

Alderman, D. and P. W. Holland. 1981. Item performance across native language groups on the Test of English as a Foreign Language. Princeton: Educational Testing Service.

Alderson, J. C. and A. Urquhart. 1985a. The effect of students' academic discipline on their performance on ESP reading tests. *Language Testing* 2: 192–204.

Alderson, J. C. and A. Urquhart. 1985b. This test is unfair: I'm not an economist. In P. Hauptman, R. LeBlanc and M.B. Wesche (eds.), *Second Language Performance Testing*. Ottawa: University of Ottawa Press.

Alderson, J. C., K. Krahnke and C. Stansfield. 1987 (eds.), *Reviews of English Language Proficiency Tests*. Washington, DC: TESOL.

Alderson, J. C., C. Clapham and D. Wall. 1995. *Language Test Construction and Evaluation*. Cambridge, UK: Cambridge University Press.

American Psychological Association 1954. *Technical Recommendations for Psychological Tests and Diagnostic Techniques*. Washington, DC. Author.

American Psychological Association 1966. *Standards for Educational and Psychological Tests and Manuals*. Washington, DC. Author.

American Psychological Association 1974. *Standards for Educational and Psychological Tests*. Washington, DC. Author.

American Psychological Association 1985. *Standards for Educational and Psychological Testing*. Washington, DC. Author.

American Psychological Association 1999. Standards for Educational and Psychological Tests. Washington, DC. Author.

Angoff, W. 1988. Validity: an evolving concept. In H. Wainer and H. Braun (eds.), *Test Validity* (pp. 19–32). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bachman, L. 1990. *Fundamental Considerations in Language Testing*. Oxford, UK: Oxford University Press.

Bachman, L., F. Davidson, K. Ryan and I-C. Choi. 1995. *An Investigation into the Comparability of Two Tests of English as a Foreign Language*. Cambridge, UK: Cambridge University Press.

Bachman, L. and A. Palmer. 1996. *Language Testing in Practice*. Oxford, UK: Oxford University Press.

Barbour, I. 1993. *Ethics in an Age of Technology*. San Francisco, CA: Harper Collins.

Baron, M., P. Pettit and M. Slote (eds.). 1997. *Three Methods of Ethics*. Malden, MA: Blackwell.

Brown, A. 1993. The role of test-taker feedback in the test development process: Test takers' reactions to a rape-mediated test of proficiency in spoken Japanese. *Language Testing* 10: 3, 277–304.

Brown, J. D. 1996. *Testing in Language Programs*. Upper Saddle River, NJ: Prentice-Hall Regents.

Camilli, G. and L. Shepard. 1994. *Methods for Identifying Biased Test Items.* Thousand Oaks, CA: Sage.

Canale, M. 1988. The measurement of communicative competence. *Annual Review of Applied Linguistics* 8: 67–84.

Chen, Z. and G. Henning. 1985. Linguistic and cultural bias in language proficiency tests. *Language Testing* 2: 155–163.

Cizek, G. (ed.). 2001. *Setting Performance Standards.* Mahwah, NJ: Lawrence Erlbaum Associates.

Clapham, C. 1996. *The Development of IELTS.* Cambridge, UK: Cambridge University Press.

Clapham, C. 1998. The effect of language proficiency and background knowledge on EAP students' reading comprehension. In A. J. Kunnan (ed.), *Validation in Language Assessment* (pp. 141–168). Mahwah, NJ: Lawrence Erlbaum Associates.

Code of Fair Testing Practices in Education. 1988. Washington, DC: Joint Committee on Testing Practices. Author.

Corson, D. 1997. Critical realism: an emancipatory philosophy for applied linguistics? *Applied Linguistics*, 18: 2, 166–188.

Crisp, R. and M. Slote (eds.). 1997. *Virtue Ethics.* Oxford, UK: Oxford University Press.

Cumming, A. 1994. Does language assessment facilitate recent immigrants' participation in Canadian society? TESL Canada Journal 2: 2, 117–133.

Davies, A. (ed.). 1968. *Language Testing Symposium: A Psycholinguistic Approach.* Oxford, UK: Oxford University Press.

Davies, A. 1977. *The Edinburgh Course in Applied Linguistics, Vol. 4.* London, UK: Oxford University Press.

Davies, A. (Guest ed.). 1997a. Ethics in language testing. *Language Testing* 14: 3.

Davies, A. 1997b. Demands of being professional in language testing. *Language Testing* 14: 3, 328–339.

Elder, C. 1996. What does test bias have to do with fairness? *Language Testing* 14: 261–277.

Educational Testing Service 1997. *Program Research Review.* Princeton, NJ: Author.

Frankena, W. 1973. Ethics, 2nd ed. Saddle River, NJ: Prentice-Hall.

Genesee, F. and J. Upshur 1996. *Classroom-based Evaluation in Second Language Education.* Cambridge, UK: Cambridge University Press.

Ginther, A. and J. Stevens 1998. Language background, ethnicity, and the internal construct validity of the Advanced Placement Spanish language examination. In A. J. Kunnan (ed.), *Validation in Language Assessment* (pp. 169–194). Mahwah, NJ: Lawrence Erlbaum Associates.

Groot, P. 1990. Language testing in research and education: The need for standards. *AILA Review* 7: 9–23.

Hale, G. 1998. Student major field and text content: Interactive effects on reading comprehension in the TOEFL. *Language Testing* 5: 49–61.

Hamp-Lyons, L. 1997a. Washback, impact and validity: ethical concerns. *Language Testing* 14:3, 295–303.

Hamp-Lyons, L. 1997b. Ethics in language testing. In C. Clapham and D. Corson (eds.), *Encyclopedia of Language and Education.* (Volume 7, Language Testing and Assessment) (pp. 323–333). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Harris, D. 1969. *Testing English as a Second Language.* New York, NY: McGraw-Hill.

Henning, G. 1987. *A Guide to Language Testing.* Cambridge, MA: Newbury House.

Holland, P. and H. Wainer (eds.). 1993. *Differential Item Functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Hughes, A. 1989. *Testing for Language Teachers.* Cambridge, UK: Cambridge University Press.

Impara, J. and B. Plake (eds.). 1998. *13th Mental Measurements Yearbook.* Lincoln, NE: The Buros Institute of Mental Measurements, University of Nebraska-Lincoln.

International English Language Testing System. 1999. *Research Reports.* Cambridge, UK: UCLES.

Kane, M. 1992. An argument-based approach to validity. *Psychological Bulletin* 112: 527–535.

Keyser, D. and R. Sweetland (eds.). 1994. *Test Critiques 10.* Austin, TX: Pro-ed.

Kim. J-O. and C. Mueller 1978. *Introduction to Factor Analysis.* Newbury Park, CA: Sage.

Kunnan, A. J. 1990. DIF in native language and gender groups in an ESL placement test. *TESOL Quarterly* 24: 741–746.

Kunnan, A. J. 1995. *Test Taker Characteristics and Test Performance: A Structural Modelling Approach.* Cambridge, UK: Cambridge University Press.

Kunnan, A. J. 2000. Fairness and justice for all. In A. J. Kunnan (ed.), *Fairness and Validation in Language Assessment* (pp. 1–14). Cambridge, UK: Cambridge University Press.

Lado, R. 1961. *Language Testing.* London, UK: Longman.

Lynch, B. 1997. In search of the ethical test. *Language Testing* 14: 3, 315–327.

Messick, S. 1989. Validity. In R. Linn (ed.), *Educational Measurement* (pp. 13–103). London: Macmillan.

Norton, B. 1997. Accountability in language testing. In C. Clapham and D. Corson (eds.), *Encyclopedia of Language and Education.* (Volume 7, Language Testing and Assessment) (pp. 313–322). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Norton, B. and P. Stein. 1998. Why the 'monkeys passage' bombed: tests, genres, and teaching. In A. J. Kunnan (ed.), *Validation in Language Assessment.* (pp. 231–249). Mahwah, NJ: Lawrence Erlbaum Associates.

Oltman, P., J. Stricker and T. Barrows. 1988. Native language, English proficiency and the structure of the TOEFL. TOEFL Research Report 27. Princeton, NJ: Educational Testing Service.

Pojman, L. 1999. Ethics, 3rd ed. Belmont, CA: Wadsworth Publishing Co.

Rawls, J. 1971. *A Theory of Justice*. Cambridge, MA: Belknap Press of Harvard University Press.

Ross, W. 1930. *The Right and the Good.* Oxford, UK: Oxford University Press.

Ryan, K. and L. F. Bachman. 1992. Differential item functioning on two tests of EFL proficiency. *Language Testing* 9:1, 12–29.

Sen, A. and B. Williams. 1982. (eds.). *Utilitarianism and Beyond.* Cambridge, UK: Cambridge University Press.

Shohamy, E. 1997. Testing methods, testing consequences: are they ethical? Are they fair? *Language Testing* 14: 340–349.

Smart, J. and B. Williams. 1973. *Utilitarianism; For and Against.* Cambridge, UK: Cambridge University Press.

Spolsky, B. 1981. Some ethical questions about language testing. In C. Klein-Braley and D. K. Stevenson (eds.), *Practice and Problems in Language Testing 1* (pp. 5–21). Frankfurt, Germany: Verlag Peter Lang.

Spolsky, B. 1995. *Measured Words*. Oxford, UK: Oxford University Press.

Spolsky, B. 1997. The ethics of gatekeeping tests: what have we learned in a hundred years? *Language Testing* 14: 3, 242–247.

Stansfield, C. 1993. Ethics, standards, and professionalism in language testing. *Issues in Applied Linguistics* 4: 2, 189–206.

Stevenson, D. K. 1981. Language testing and academic accountability: on redefining the role of language testing in language teaching. *International Review of Applied Linguistics* 19: 15–30.

Stigler, S. 1986. *The History of Statistics*. Cambridge, MA: Belknap Press of Harvard University Press.

Taylor, C., J. Jamieson, D. Eignor and I. Kirsch. 1998. The relationship between computer familiarity and performance on computer-based TOEFL tests tasks. *TOEFL Research Report No. 61.* Princeton, NJ: Educational Testing Research.

University of Michigan English Language Institute 1996. *MELAB Technical Manual.* Ann Arbor, MI: University of Michigan Press. Author.

Valdés, G. and R. Figueroa. 1994. *Bilingualism and Testing: A Special Case of Bias*. Norwood, NJ: Lawrence Erlbaum Associates.

Wall, D. and Alderson, C. 1993. Examining washback: the Sri Lankan impact study. *Language Testing* 10: 41–70.

Willingham, W. W. and N. Cole. 1997. *Gender and Fair Assessment.* Mahwah, NJ: Lawrence Erlbaum Associates.

Willingham, W.W. 1999. *A systemic view of test fairness.* In S. Messick (ed.), *Assessment in Higher Education: Issues of Access, Quality, Student Development, and Public Policy* (pp. 213–242). Mahwah, NJ: Lawrence Erlbaum Associates.

Zeidner, M. 1986. Are English language aptitude tests biased towards culturally different minority groups? Some Israeli findings. *Language Testing* 3: 80–95.

Zeidner, M. 1987. A comparison of ethnic, sex and age biases in the predictive validity of English language aptitude tests. Some Israeli data. *Language Testing* 4: 55–71.

Appendix 1

Test fairness framework

Table 1: Test fairness framework

Main quality	Main focus
1. Validity	
Content representativeness/coverage ➡	Representativeness of items, tasks, topics
Construct or theory-based validity ➡	Representation of construct/underlying trait
Criterion-related validity ➡	Test score comparison with external criteria
Reliability ➡	Stability, Alternate form, Inter-rater and Internal consistency
2. Absence of bias	
Offensive content or language ➡	Stereotypes of population groups
Unfair penalisation ➡	Content bias based on test takers' background
Disparate impact and standard setting ➡	DIF in terms of test performance; criterion setting and selection decisions
3. Access	
Educational ➡	Opportunity to learn
Financial ➡	Comparable affordability
Geographical ➡	Optimum location and distance
Personal ➡	Accommodations for test takers with disabilities
Equipment and conditions ➡	Appropriate familiarity
4. Administration	
Physical setting ➡	Optimum physical settings
Uniformity and security ➡	Uniformity and security
5. Social consequences	
Washback ➡	Desirable effects on instruction
Remedies ➡	Re-scoring, re-evaluation; legal remedies

Appendix 2

Notes

1 Angoff (1988) notes that this shift is a significant change.
2 See this document for a full listing of titles and abstracts of research studies from 1960 to 1996 for TOEFL as well as other tests such as SAT, GRE, LSAT and GMAT.
3 The FCE stands for First Certificate in English, CPE for Certificate of Proficiency in English and IELTS for International English Language Testing Service.
4 Another organisation, the Association of Language Testers of Europe (ALTE), of which UCLES is a member, has a *Code of Practice* that closely resembles the Code of Fair Testing Practices in Education (1988). However, there are no published test evaluation reports that systematically apply the *Code*.
5 MELAB stands for the Michigan English Language Assessment Battery.
6 Recent reviews in *Language Testing* of the MELAB, the TSE, the APIEL and the TOEFL CBT have begun to discuss fairness (in a limited way) along with traditional qualities such as validity and reliability.
7 This uniformity is probably also due to the way in which MMY editors prefer to conceptualise and organise reviews under headings, such as description, features, development, administration, validity, reliability and summary.
8 For DIF methodology, see Holland and Wainer (1993) and Camilli and Shepard (1994).
9 For arguments for and against utilitarianism, see Smart and Williams (1973) and Sen and Williams (1982).
10 Bentham, the classical utilitarian, invented a scheme to measure pleasure and pain called the Hedonic calculus, which registered seven aspects of a pleasurable or painful experience: intensity, duration, certainty, nearness, fruitfulness, purity and extent. According to this scheme, summing up the amounts of pleasure and pain for sets of acts and then comparing the scores could provide information as to which acts were desirable.
11 See Ross (1930) and Rawls (1971) for discussions of this system.
12 See Crisp and Slote (1997) and Baron, Pettit and Slote (1997) for discussions of virtue-based ethics. Non-secular ethics such as religion-based ethics, non-Western ethics such as African ethics, and feminist ethics are other ethical systems that may be appropriate to consider in different contexts.
13 See Rawls' (1971) *A Theory of Justice* for a clear exposition of why it is necessary to have an effective sense of justice in a well-ordered society.
14 These principles are articulated in such a way that they complement each other and if there is a situation where the two principles are in conflict, Principle 1 (The Principle of Justice) will have overriding authority. Further, the sub-principles are only explications of the principles and do not have any authority on their own.
15 The authors of the document also acknowledge that many additional interpretations of the term 'fairness' may be found in the technical testing and the popular literature.
16 There is substantial literature that is relevant to bias and DIF in language testing. For empirical studies, see Alderman and Holland (1981), Chen and Henning (1985), Zeidner (1986, 1987), Oltman *et al.* (1988), Kunnan (1990), Ryan and Bachman (1992).

17 For standard setting, the concept and practice, see numerous papers in Cizek (2001).

18 In the US, Title VII of the Civil Rights Act of 1964 provides remedies for persons who feel they are discriminated against owing to their gender, race/ethnicity, native language, national original, and so on. The Family and Education Rights and Privacy Act of 1974 provides for the right to inspect records such as tests and the right to privacy limiting official school records only to those who have legitimate educational needs. The Individuals with Disabilities Education Amendments Act of 1991 and the Rehabilitation Act of 1973 provide for the right of parental involvement and the right to fairness in testing. Finally, the Americans with Disabilities Act of 1990 provides for the right to accommodated testing. These Acts have been used broadly to challenge tests and testing practices in court.