Language Assessment From a Wider Context

Antony John Kunnan
California State University, Los Angeles

INTRODUCTION

Most individuals who have been required to take language tests know that tests are part of the sociopolitical setup of a community. Yet, language assessment is characterized as a field that is primarily concerned with the psychometric qualities of tests and one in which test developers/researchers ignore the socioeconomic-political issues that are critically part of tests and testing practice. It is not that these perspectives have not been known earlier. In fact, about two decades ago, Cronbach (1984) pointed out that “testing abilities has always been intended as an impartial way to perform a political function—that of determining who gets what” (p. 5). Bachman (1990) also succinctly stated that “tests are not developed and used in a value-free psychometric test tube; they are virtually always intended to serve the needs of an educational system or of society at large” (p. 279). In addition, Hanson (1993) characterized the intentions and social benefits of testing:

Most tests are intended to contribute to social ends that are generally reckoned as beneficial, such as equal opportunity, honesty, law-abiding behavior, acquisition of knowledge and skills, identification and development of individual interests and talents, knowledge, and skills for the mutual benefit of the individual and society. (p. 6)

Examining the intentions and social benefits of language testing in the 20th century, Spolsky’s (1995) pioneering and eye-opening treatise entitled Measured Words was critical of the whole enterprise:

Since the days of World War I, psychometric principles and practices have come to dominate the testing of foreign language proficiency, and a movement that initially blossomed in the United States has spread throughout the world. As long as testing was confined to helping students learn or to determining the qualifications of individuals seeking employment, there was a strong ethical case to be made for it, as the end justifies the means. But, from its beginnings, testing has been exploited also as a method of control and power—as a way to select, to motivate, to punish. The so-called objective test, by
virtue of its claim of scientific backing for its impartiality, and especially when it operates under the aegis and with the efficiency of big business, is even more brutally effective in exercising this authority. (p. 1)

He also called for a different approach to understanding the motivations and impact of tests.

It is only by taking full account of the institutional or political context that one can appreciate how the psychometric controversies have distracted attention from more serious social (or antisocial) motivations and impact. (p. 1)

Having set up the framework, Spolsky (1995) convincingly showed the influence of personal, institutional, political, and economic policies on the development of language testing in the United States (particularly in the development of the Test of English as a Foreign Language at Educational Testing Service, Princeton) and in the United Kingdom (at the University of Cambridge Local Examinations Syndicate), often overriding appropriate educational and ethical practice.

Using a similar framework, Shohamy (2001) found compelling evidence in three case studies (a reading test, an Arabic test, and an English test) that the “power of tests and their detrimental forces” can alter the intentions of those who introduce tests. McNamara (1998) also noted that tests are used as a disciplinary tool by political institutions for example, as an arm of policy reform and control in fields as broad as education, vocational training, immigration, and citizenship.

This chapter follows these researchers in spirit. I examine tests and testing practice from a wide context in order to more fully determine whether and how these tests are beneficial or detrimental to society. The wide “context”, constructed from reflection and research in the last two decades, includes the political and economic, the educational, social and cultural, the technological and infrastructure, and the legal and ethical.

TEST VALIDITY AND TEST FAIRNESS: OPENING UP THE “CONTEXT”

In the late 1980s, Messick (1989) revolutionized test validity discussions by arguing for a unified view of validity. Specifically, he asserted that validity should be considered as a unified concept (in contrast to the three traditional validity types: content validity, predictive and concurrent criterion-related validity, and construct validity developed and applied in earlier decades) with a superordinate role for construct validity. He postulated that test validity refers to the “appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores” (p.8) and that the unified validity framework could be constructed “by distinguishing two interconnected facets of the unified validity concept. One facet is the source of justification of the testing, being based on appraisal of either evidence or consequence. The other facet is the function or the outcome of the testing, being either interpretation or use” (Messick, 1989, p. 20)

In this view of validity, Messick also explicitly advanced a critical role for value implications and social consequences, particularly evaluation of intended and unintended social consequences of test interpretation and use, as part of test validity. This was the first time that values implications and social consequences were brought from the back room (where test developers had conveniently ignored them) and included as part of test validity. This view has now been instantiated in the 1999 Standards (AERA, APA, NCME, 1999). The examination of the social value of tests as well as their unanticipated consequences or side effects, especially if such effects were traceable to sources of invalidity of test score interpretation, received support from this view. Many researchers welcomed this significant development as a possible sign of a new beginning
in a hitherto psychometrically driven field and such discussions are widespread today (see paper in Cumming & Berwick, 1996; Kunnan, 1998; and Chapelle, 1999). Messick's view of test validity also triggered the reworking of the role of test reliability and has led many researchers to argue that test reliability evidence should be used as evidence that contributes to test validity (Chapelle, 1999; Kunnan, 2000).

Similarly, after Messick's view was widely circulated, the importance of fairness has also been understood better. The 1999 Standards acknowledged the critical importance of fairness as a goal in testing by devoting a whole chapter to fairness for the first time. Its vision is as follows:

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. It would consider the technical properties of tests, the ways test results are reported, and the factors that are validly or erroneously thought to account for patterns of test performance for groups and individuals. A comprehensive analysis would also examine the regulations, statutes, and case law that govern test use and the remedies for harmful practices. (p. 73; emphasis added)

Further more, the Standards specifically defined three categories of fairness: fairness as lack of bias, fairness as equitable treatment in the testing process, and fairness as equality in outcomes of testing.

Willingham and Cole (1997) offered another perspective, but their focus was noticeably different: Their focus was on comparable validity, the validity of score interpretations for all individuals and groups. Their three criteria for fair tests were comparable opportunity for test-takers to demonstrate relevant proficiency, comparable testing tasks and scores, and comparable treatment of test-takers in test interpretations and use. Incorporating these two definitions and Messick's interest in values and social consequences through his unified validity framework, Kunnan (2004) proposed a fairness framework that positions fairness as the ultimate goal in testing. The main qualities of the framework are construct validity, absence of bias, access, administration, and social consequences. These concepts inform the "Wider Context" concept.

THE WIDER CONTEXT

The wider context of tests and testing practice refers mainly to the collection of traditions, histories, customs, professional practices, and academic, social, and political institutions of a community. This collection can then be identified loosely as the political and economic, the educational, social and cultural, the technological and infrastructure, and the legal and ethical contexts of a community in which a test operates (Figure 43.1). Other contexts that may also play a role in a community but which are not explicitly shown in Fig. 43.1 include race and ethnicity, gender, class, caste, religion, sexual orientation, entertainment, and so on. As shown in the figure, the main contexts surround, overlap, and enmesh each other and it is into this milieu that a test is thrust when it is commissioned. It is then developed, administered, scored, reported, and researched and decisions are made based on test scores for this community. This conceptualization also implies that we need to use this wider context in debating and evaluating tests and testing practice but does not exclude any of the technical aspects of language testing practice, such as validity, reliability, standard-setting, and the like from being used in such evaluations.

The Political and Economic Context

The political and economic context of language assessment has not been—until very recently—overtly acknowledged, and relative silence on this front has contributed
to the notion that language testing is an apolitical discipline (see Spolsky, 1995; McNamara, 1998; Shohamy, 2001, for notable exceptions). An examination of school-level testing and testing for immigration and citizenship will be discussed along these lines.

**The Politicization of School-Level Testing**

In the United States, there has recently been a huge growth in publications and public discussions on school-level standardized testing, so much so that electoral campaigns and platforms have become a key stage for such debates. Elected officials (such as the president, state governors, mayors, and school superintendents) often take the first opportunity to underscore their desire to impose standardized testing on schools in their jurisdiction in the guise of public accountability.

Large-scale standardized testing is now the main component that drives the concept of public school accountability. This is largely because until recently, most school districts did not require tests for grade-level promotion or high school graduation. Instead, schools depended on teacher grades for courses and students were promoted to the next grade automatically. College-bound students took additional courses and standardized tests such as the SAT for admission to colleges and universities. The official accountability argument goes as follows: If parents knew how well their children were doing, then educators, policymakers, and the public would know how well their schools were doing. Financial incentives would then be offered to schools that have met or exceeded their goals or punitive action would be taken against schools that do poorly.

President Bush’s education agenda added negatively to this state of affairs when he signed the *No Child Left Behind (NCLB) Act of 2001*, which requires all states to introduce testing in science, mathematics, and (English language) reading for Grades 3 through 8. It is based on four basic principles: stronger accountability for results, increased flexibility and local control, expanded options for parents, and an emphasis on teaching methods that have been proven to work. In terms of increased accountability, the Act requires states to implement statewide accountability systems covering all public schools and students on state standards in reading and mathematics, annual testing for all students in Grades 3–8, and annual statewide progress objectives ensuring that all groups of students reach proficiency within 12 years. Results of such
tests are required to be broken out by poverty, race, ethnicity, disability, and limited English proficiency to ensure that no group is left behind. School districts and schools that fail to make adequate yearly progress (AYP) toward statewide proficiency goals will, over time, be subject to improvement, corrective action, and restructuring measures. Schools that meet or exceed AYP objectives or close achievement gaps will be eligible for State Academic Achievement Awards. This overtly stated goal masks the Bush administration’s related concept of awarding parents tuition fees (known as the School Voucher Program) if they choose to remove their children from failing public schools and place them in private schools.

Whereas there is general support for the accountability argument from educators and parent groups, (which is driven by dissatisfaction with instructional and grading practices and social promotion through grade levels), the mechanism of using standardized testing to achieve this end has been challenged by education groups, including civil rights organizations, as well as individuals. For example, according to Heubert and Hauser (1999), “for more than 20 years the National Association for the Advancement of Colored People has called the use of testing as a sole criterion for the nonpromotion of students and the use of competency testing for high school graduation ‘another way of blaming the student victim’” (p. 45). The Mexican American Legal Defense and Educational Fund has filed a lawsuit in Texas against the use of the Texas Testing of Academic Skills, which claims that, in the words of Heubert and Hauser (1999), “the test denies diplomas to students without sufficient proof that the test does not correspond to what is actually taught in schools in many minority communities” (p. 46). Similarly, the National Association of Bilingual Education argues, according to Heubert and Hauser (1999) “students should be assessed with appropriate, performance-based tests, and English language learners should not be assessed with tests that are inappropriate at their level of language competency” (p. 46). At the individual level, Linn, Baker, and Betebenner (2002) present arguments that the states’ content standards, the rigor of their tests, and the stringency of their performance standards vary greatly, thus posing substantial challenges to schools, school districts, and states.

Overall, there is a disjuncture between elected officials and opinion polls, on the one hand, and major educational groups and academic researchers regarding school-level testing on the other hand. However, school-level tests continue to be used as policy instruments by elected officials. Whatever direction this policy takes in the near future, test use will continue to be a highly politicized issue, as ultimately whether and how tests are used is largely decided by elected officials in school boards, governments, and legislatures, and finally in courts; in most cases without much regard to opinions from educators and academic researchers. As Cronbach (1980) put it, “the public intends to judge specific test uses for itself, directly or through courts and legislatures” (p. 100).

**Testing for Immigration and Citizenship: Political and Economic Gatekeeping**

Political and economic interests have been at the center of gatekeeping in the high-stakes arena of immigration and citizenship. In most countries, language abilities (in the official language of the country) have been required of potential immigrants. Often, the overt goal of assessing language proficiency and the associated intention of providing free language instruction is not the real reason for the testing policy. The real objective may be “racial exclusion,” as in the case of the dictation test in Australia and in the case of intelligence tests in the United States in the early 20th century.

In the case of Australia, language tests were used as part of the immigration requirement as early as 1901 as part of the “White Australia” policy. The first test was a dictation test in a European language (or any “prescribed language”) to a potential
immigrant in such a way that the immigrant (if not from the British Isles) could be excluded. As a result of rigorous application of this policy from 1902 to 1946, only 125,000 members of “the alien races” (Asians and “colored”) were admitted to the country. In 1956, the dictation test given in any European language was changed to an English dictation test. In the 1990s, first the ACCESS and then the STEP tests were introduced to assess English language proficiency of professionals prior to registration or immigration of those seeking permanent residence. However, standard setting was influenced by immigration policy that was intended to be benign but sometimes indefensible in practice. According to Hawthorne (1997), the standard set on the ACCESS test restricted immigration in terms of English language ability but the standard on the STEP test was set in such a way as to enable Chinese students already in Australia (after the events of Tiananmen Square) to continue their stay: “The case of the STEP test offers a dramatic illustration of the increasing use of language testing by Australian authorities to achieve political purposes... The federal government was able to impose control over a politically volatile situation; the Australian legal system was cleared of an unmanageably large backlog of refugee applications; and young PRC, Pakistani, Sri Lankan, and other asylum seekers were transformed into a relatively educated and acculturated skilled migrant intake—commencing their legal permanent residence in Australia supported by a functional basis of English” (pp. 257–258).

Rather similarly, the use of intelligence tests for immigration in the 1920s had a “racial exclusion” objective in The United States. The earliest language testing was the English Competence examination prepared by the College Entrance Examination Board in 1930. According to Spolsky (1995), this test was intended to deal with a loophole in the Immigration Act of 1924: “Carl Brigham gave evidence in Congress on the deleterious effects of permitting non-Nordic immigrants to ‘contaminate the American gene pool’” (p. 55). Spolsky (1995) cites Brigham’s racist intentions: “We must frankly admit the undesirable results which would ensue from a cross between the Nordic in this country with the Alpine Slav, with the degenerated hybrid Mediterranean, or with the Negro, or from the promiscuous intermingling of all four types... Immigration should not only be restrictive but highly selective” (p. 37). In addition, according to Spolsky (1995), the nexus between the government and testing agencies was clear: “It is important to note that the purpose of this test... was in fact political and restrictive, namely to enforce an immigration act intended to close the gates to ‘undesirables.’ A well-established testing agency wrote it at government request” (p. 59).

More recently, the political and economic interests of New Zealand government were evident in the way it introduced the IELTS in 1995. As Read (2001) notes, “the test which would determine whether applicants in the target categories had a minimum level of English proficiency, defined as Level 5 in the General module. Principal applicants were required to pass the test in all four skills before arriving in New Zealand. Other members of the family (or ‘nonprincipal applicants’) aged 16 and over were also expected to achieve the same level. However, they could be admitted to the country without doing so, upon payment of a fee of NZ$20,000 each. The fee was refundable as follows: a full refund would be given if IELTS Level 5 was achieved within three months of arrival; the sum of NZ$14,000 would be refunded if the required level was reached from three to twelve months after arrival. After that, the full NZ$20,000 was to be retained by the government” (p. 192). Read (2001) goes on to state that the government realized that this high fee was having a negative effect on potential business investor applicants, [and] as an intermediate measure the level of attainment for refund was lowered to Level 4. Later in 1998, the NZ$20,000 fee was abolished and was replaced with a fee ranging from NZ$1,700 to NZ$6,650, which was to entitle the applicants to ESL tuition for a 3-year period.

Many countries also have language tests for citizenship. In the United States, in order to become a citizen, one of the requirements is that the applicant “must
understand English and be able to read, write, and speak ordinary English unless the applicant is 50 years or older and have lived in the county for 20 years\(^5\) (Immigration and Naturalization Service [INS])\(^6\). Although this requirement is stated clearly, what is unclear is how it is operationalized, as there does not seem to be a standard testing protocol or procedure. The INS examiners who conduct the final interview also conduct the civics test and the English test, along with other duties such as examining documents. Unlike the civics part, which is made up of a body of knowledge that is described in the citizenship study materials, the English test is not described in terms of a proficiency level or can-do statements and is therefore left to the interpretation of the examiner, allowing obvious variations in the test. Anecdotal information collected from citizenship applicants indicates the test may consist of any of the following: The applicant may be asked to write a sentence in English (such as “Today is a beautiful day”), read the civics questions aloud and answer the questions in English (“Who was the first president of the United States?”), or answer questions on everyday matters in English (“Where do you live?”)\(^7\). Such variation in any test would be cause for concern; in this high-stakes test, it is a vexing problem that needs immediate attention.

In the 1990s, new language tests have become part of the requirements for application for citizenship in many countries such as Estonia, Germany, Latvia, and Lithuania. According to Shohamy (2001), the test of Latvian required of all residents including Russians, many of whom have lived their entire lives in Russian-speaking communities in Latvia, is an example of a test used to discriminate against minority ethnic groups: “...in 1996, a report showed a drastic decline in the number of Russians residing in Latvia...While the language test may not be the only reason for the decline, it clearly contributed significantly to ethnic cleansing” (p. 89).

From these examples, it is imperative for language testers to examine the social consequences of such policies, particularly in terms of what would happen to the applicants who do not pass the test and therefore cannot become citizens. Would they have to continue to live in such countries as second-class citizens because of their inability to achieve an acceptable level of language ability or return to their “homelands”? More fundamental questions could be asked of immigration tests: What is the purpose of such a test? Is the purpose to encourage applicants to learn the local language or identify their deficiencies so that they can take the language classes they need? Or, is the purpose to expressly deny rights and powers to residents who have not mastered the local language to an acceptable level and by extension have not assimilated to the local community and country? Similar questions have been raised in the context of testing immigrants in Canada (Cumming, 1997).

**The Educational, Social, and Cultural Context**

Standardized language tests are common in the educational, social, and cultural contexts. They are used for a variety of educational and career-related purposes that include competency, admission, and employment. Key concerns that have been raised about these tests are primarily regarding their educational and social consequence: test bias and washback.

**Standardized Tests and Washback**

Standardized tests have become most popular in U.S. education in. Scores on the SAT that assess verbal and quantitative ability and/or SAT-II (subject areas) or the ACT are required for admission to U.S. colleges\(^10\). Similarly, students who are non-native speakers of English and who wish to apply for admission to U.S. and Canadian colleges and universities are required to demonstrate their English language ability by taking and providing scores on the Test of English as a Foreign Language (TOEFL)
or the International English Language Testing System (IELTS). At the graduate level, applicants are generally required to demonstrate their preparedness for study by providing scores on graduate-level aptitude tests (that assess English language reading and writing ability in addition to quantitative skills) such as the Graduate Record Examination or on professional tests such as the Graduate Management Test.

In keeping with the accountability concept, the state of California introduced the California High School Exit Examination (CAHSEE) in English from 2003–2004. According to the State Education Code, the purpose of the CAHSEE is to ensure that students who graduate from high school can demonstrate grade-level competency in the state content standards for reading, writing, and mathematics. The exam has two parts, English-language arts and mathematics, and students of the class of 2006 must pass both portions of the test to receive their high school diploma. Although the independent evaluation of the test reports that the test development, administration, scoring, and reporting is flawless, the report documents through one table that teachers spent 45 hours on test preparation and in one short paragraph that a small DIF study was conducted for Hispanic test-takers.12

In 1998, anticipating the national accountability concept, California introduced standardized tests that assesses students from Grades 2 to 11 in a variety of subjects including English reading and writing, mathematics, and science. The impact of this test has generally been negative particularly in terms of washback: Teacher anecdotes have indicated that they have to “teach to the test” for about two months prior to the test and that the curriculum has in general become “test-driven” with less time devoted to activities that are not part of the test. Furthermore, there have been complaints that the test is not aligned to the stated curriculum, a clear problem of content representativeness/coverage. This is a result of many factors, including the situation that different school districts have slightly different curricula and different school districts have slightly different timelines for completion of topics, units, and concepts. It is clear, therefore, that more systematic washback studies need to be conducted although the limitations and dangerous consequences of such standardized tests have been regularly raised13. The value of these high-stakes educational tests to California society depends on construct validity, and the absence of bias and positive washback of the tests on instruction.

Achieving Fairness Through Test Bias Studies

Fair access to education and employment has required test bias investigations on standardized tests. Although such investigations are not always mandatory, they are advisable in multicultural, multilingual test-taking populations where a test may be considered biased when it produces systematic differential performance among test-takers from the same ability group but from different subgroups of interest (such as age, academic major, gender, race and ethnicity, nationality, native language, religion, and test-takers with disability). Such systematic differential performance may be due to the presence of construct-irrelevant test characteristics in a test (or in test items) or relevant secondary item characteristics. Irrelevant item characteristics may be found in different components of a test: language variety, directions, content, response process, administration, and scoring, reporting, interpretation toward a particular subgroup. Thus, differences in test performance for a designated subgroup of interest (DSI) might result in differences in meaning of test scores such that the validity of the test scores would be seriously in doubt. In summary, test bias could occur when two groups of test-takers with equal ability show a differential probability of a correct response. Furthermore, as test bias investigations are based on comparable test-takers matched with respect to the primary construct the test is measuring, a large difference in performance may mean the test is measuring additional constructs differently across the
groups. The additional construct(s) could be either intended secondary constructs or
nuisance constructs that have crept into the test.

According to the 1999 Standards, there are two sources of bias: content-related
and response-related. Content-related bias is related to the following aspects of
testing:

1. *Content representativeness:* In an achievement test, it is possible there may be
less than desirable alignment between the curriculum and the test items leading to
content underrepresentativeness or construct irrelevance. A slightly different type of
content bias may also be identified in cases where the topic or subject matter or events,
places, and situations that are construct-irrelevant are different or unfamiliar to a DSI
(such as test-takers from different academic major, gender or cultural or educational
background subgroups).

2. *Language variety:* The language of the test can be a source of bias, particularly if
a DSI uses a language variety different from the language variety of a test (such as a
DSI uses British or Australian English instead of American English that is used in the
test to measure English language ability).

3. *Offensive content and language:* This type of bias refers to content or language
used in a test that is controversial or inflammatory to a DSI, demeaning or offensive
to members of a DSI, depicting members of a DSI as having stereotypical occupations
(i.e., Chinese launderer, White professionals, Black porters, Hispanic laborers, female
secretaries or nurses) or in stereotypical situations (i.e., boys as successful in math, girls
needing help with math) or material that does not provide adequate representation
of all DSI. This type of bias also includes terms and expressions that are offensive or
inappropriate, such as “he” to include male as well as female members, “man” as in
policeman instead of police officer, “fireman” instead of firefighters, and “disabled
workers” instead of workers with disability.

Response-related source of bias is related to the following aspects of testing:

1. *Response-format:* This type of bias refers to the format used in a test such as the
item format as well as the response format (essay writing, short-answer, multiple-
choice, true-false, etc.). It also includes test directions, scoring rubrics, or test-taking
procedure that may be ambiguous or confusing to a particular DSI.

2. *Differential performance:* This type of bias refers to differences in performance on
a test between two subgroups (for example, age, gender, race/ethnicity, geographic,
or other DSI).\textsuperscript{13}

3. *Standard setting:* When a test is used for standard setting (that is, selection and
prediction), evidence is gathered to check whether the relationship between the test
score and the criterion score (used for standard setting) is biased for a DSI.

Many standardized language tests have been the subject of test bias studies. Most
of the attention has been on the following areas: language variety, language content,
response format, differential performance, and standard setting (see Kunnan, 2000,
for a list of studies). Such investigations have helped test developers and researchers
understand the role of social and cultural factors in the performance of different test-
taker groups although the limited role of such studies in terms of fairness has also
been pointed out (Camilli & Shepard, 1997; Kunnan, 2000).

**The Technology and Infrastructure Context**

The importance of this context has become clearer in the last decade than ever before
due to the rapid use of high technology in testing such as computers and the Internet.
A key concern with the use of technology often articulated is whether test-takers have access to and knowledge of the technology necessary for success on a test.

**Technology and Access**

In low-technology-use areas, where machine scoring has replaced human scoring, it is likely that this is coupled with the promotion of multiple-choice test items and the use of machine-scorable cards that enable machine scoring (such as most U.S. based tests). While recording responses to test items on a machine-scorable card might be easy for test-takers who are used to this format, test-takers who are not used to it might find it rather confusing and unnecessarily tricky from two points of view: the use of multiple choices for each test item and the recording of responses. Thus, it is critical that test developers are aware of how familiar test-taking groups (in their targeted population) are with multiple-choice test items and whether they are able to record their responses.

Similarly, in the case of tape-mediated tests (such as the Test of Spoken English or the Simulated Oral Proficiency Tests), talking into an audio tape recorder in response to an unseen voice that asks for responses is another problematic test format. Test-takers who are unfamiliar with this format will have difficulty dealing with it (especially starting and completing responses within the time allocated for each response). Furthermore, from a social and cultural point of view, it may seem inappropriate to particular test-taking groups to talk into a tape recorder, especially in tests that claim to assess communicative language ability.

In high-technology-use areas, computers may be used in all aspects of test development, administration, scoring, and reporting. In test administrations using computers (such as the computer-based TOEFL), a test-taker would receive test items on the computer screen that have been either based on item difficulty and person ability estimates (if the test section is a computer-adaptive test) or based on a set order of items (as in a computer-based test or a paper-and-pencil test). Either way, the test-taker would need to have the requisite computer keyboarding and mouse movement skills (clicking, scrolling, highlighting, etc.) in order to read the test items and to record the answers. This calls into question the issue of test access if the test is required of test-takers in places where such computer skills and computer-based tests are relatively new or nonexistent.

For example, the administrators of the computer-based TOEFL (which replaced the paper-and-pencil TOEFL) had to deal with this issue as the test is offered worldwide, including places in which high-technology use is new or nonexistent. As the test administrators were aware of the test access issue, a computer-familiarity study of potential TOEFL test-takers was conducted prior to launching the test (Taylor et al., 1998). This study enabled the test administrators to be confident that most test-takers who might take the computer-based TOEFL have sufficient familiarity with keyboarding and mouse skills and that those who do not could benefit from a free tutorial that test-takers could take prior to taking the test. This solution is acceptable for test-takers who come forward to take the tutorial and the test, but it would not be sufficient if the very thought of having to take a computer-familiarity tutorial and computer-based test inhibited a sizable number of test-takers in a particular area from taking the test. This example shows how the use of high technology has to be understood and managed in areas where such technology may not be commonplace among test-takers.

Other related issues also need to be considered. Cost is a concern when paper-and-pencil versions of tests are replaced with computer-based tests; geographical access is another, as a computer-based test may only be available in test centers located in cities. Furthermore, comparability of test versions (computer-based and paper-and-pencil) is also a matter of concern. Thus, if these are major concerns, then technology with computer-based testing may raise more questions than answers.
The Legal and Ethical Context

Test-takers and test score users need a remedial procedure whenever a test is shown to be in violation of established practice or a regulation. For example, if there are problems with any aspect of a test (indefensible construct validity, absence of bias, lack of access, administration, scoring, score-reporting, etc.) or if any existing regulations have been violated that has a direct impact on test-takers, remedial measures should be available to test-takers or any affected persons or agencies. Furthermore, if decisions made on test scores that are in doubt can be reversed as in nonhigh stakes tests, then any of the following remedial measures might satisfy affected test-takers: re-scoring, re-totaling, or re-taking of the test for free or a small fee. In cases when decisions are not easily reversible or test-takers are affected adversely, legal action may be the only recourse.

Legal Framework in the U.S.

In the United States, legal frameworks based on the Constitution, federal civil rights statutes, and judicial decisions apply to standardized tests (including language tests) in educational, licensure, employment, and professional arenas. Test score interpretations and decisions based on scores can be challenged on three grounds by test-takers or interested parties. They are the discrimination challenge based on the basis of test-takers' race, color, national origin, or gender; the due process challenge based on inadequate notice of test requirements or a challenge of the test's curricular validity; and the disability challenge based on legislation passed to protect Americans with disabilities.

The discrimination challenge: In terms of the discrimination challenge, three types of claims can be made to a court: (1) A test is intentionally discriminating against test-takers (who are a particular class of people) on the basis of race, color, national origin, or gender. This challenge is based on the equal protection clause in the 14th Amendment of the U.S. Constitution that forbids public employers from engaging in acts of discrimination. (2) A test preserves the effects of prior discrimination. (3) A disparate impact claim. This claim can be made when different test-taker groups receive different scores (such as female and male or from different race/ethnic groups). State and public school agencies that receive federal funding are prohibited from discriminating against students based on federal statutes.

The due process challenge: High-stakes tests that are not found to be discriminating against protected classes may still be illegal under the due process provision of the 5th and 14th Amendments of the U.S. Constitution. The claims under this provision may be either that the test-takers did not receive sufficient advance notice or adequate notice of the test or that the test-takers did not receive instruction on the test knowledge and skills (also known as curricular validity). In terms of the adequate notice provision, test score users or similar agencies are expected to provide adequate notice. This has been interpreted by courts to mean anywhere between 1 to 4 years of advance notice before a test becomes effective, so that the test-takers have adequate opportunity to learn the relevant knowledge or skills. In terms of curricular validity, there have been disagreements as to how educational agencies can demonstrate what students have been taught. Some argue that the formal written school or district curriculum can be used to match the knowledge and skills measured in the test. Others argue that it is not the formal written curriculum that should be used to check curricular validity but the instructional curriculum of the classroom. Test researchers have also made the same argument: They prefer to examine whether test-takers have had the “opportunity to learn” the knowledge and skills in the classroom rather then merely matching test knowledge and skills with a formal written curriculum.

Accommodations for test-takers with disabilities: Test accommodations for test-takers with disabilities were given a major push in the form of three pieces of legislation
as part of a political and social agenda: Section 504 of the Rehabilitation Act of 1973, the Americans with Disability Act (ADA) of 1990, and the Individuals with Disabilities Educational Act (IDEA) of 1991 and 1997. Of these, the ADA of 1990 prohibits not only discrimination against individuals with disabilities but also relates to the opportunities for individuals to obtain employment and education.\textsuperscript{17}

Following these legislative actions, the 1999 Standards and the Code of Fair Testing Practices in Education (Joint Committee on Testing Practices, 1988) recommend that test accommodations should be made available for test-takers with disabilities. In the 1999 Standards, the chapter titled “Testing Individuals with Disabilities” identifies individuals with disabilities “as persons possessing a physical, mental, or developmental impairment that substantially limits one or more of their life functions” (p. 101). And, it urges “test developers and users to become familiar with (U.S.) federal, state, and local laws, and court and administrative rulings that regulate the testing and assessment of individuals with disabilities” (p.101). U.S. federal regulations and testing guidelines, therefore, mandate that tests should not discriminate against test-takers with disabilities but must be valid measures, measuring the ability to be tested and not the disability. While this mandate for testing and test accommodations is evident, IDEA regulations and Section 504 contain an important proviso: “...except where those skills are the factors that the test purports to measure” (Burns, 1998, p. 9). This is important because in some cases the disability itself might be the focus of the test. However, neither the regulations nor the testing guidelines provide or suggest the exact test accommodations to be offered to test-takers with specific disabilities. This leaves many critical issues to be sorted out by test developers, administrators, and users, and finally, by courts.

In the past, test-takers with disabilities were identified only as persons with visual, hearing, or physical impairments. Test-takers with visual impairments were provided a Braille version of a test, test-takers with hearing impairments received a written version of test directions, and test-takers with physical impairments were offered the services of a scribe. In recent years, however, learning disabilities have gained more attention, and appropriate test accommodations (such as extended time and format change) have been designed for this group of test-takers. In fact, in 1994–1995, out of about 5 million test-takers with disabilities, a little more than 50% are learning disabled. What is critical to recognize is that as people with disabilities are heterogeneous, test accommodations need to fit the needs of the individuals if the test accommodation has to be valid as well as meaningful. The most popular type of accommodation is time accommodation: test scheduling, time management, extended time, and reduced time. A chief concern of this test accommodation is whether the validity of the test-score interpretations is altered when the extended time given is varied for test-takers with and without disabilities.

\textit{ADA-related challenges:} Prior to the passage of the ADA of 1990 and the IDEA of 1991 and 1997, legal cases filed by test-takers with disabilities or disability rights groups were based on the two challenges discussed earlier. Post-ADA legal cases have generally been related to learning disabilities (rather than physical disabilities) and this is probably due to the fact that the majority of test-takers with disabilities are now from this group. But in the various rulings, courts have not provided clear directions on many critical matters like test accommodations (such as extended time). Pithoniak and Royer (2001) state, “The results of these cases do not present a consistent picture, owing to courts’ struggling with both how to determine whether a learning disability exists and what that disability means in terms of the affected individual’s rights” (p. 63). In summary, as legal challenges in testing are relatively new, court opinions do not offer clear directions as to how to proceed with many matters, including how to avoid discriminatory testing practice, how to identify individuals with disabilities, or what test accommodations are appropriate for different physical and learning disabilities.\textsuperscript{18}
Ethics and Professional Standards: The Way Forward

With legal standards and regulations constantly under attack and reinterpretation by courts, it is obvious that legal recourse for unsound professional practice may not always help and is often too late from the test-takers' point of view. Thus, sound professional practice needs to be promoted and one way of doing this would be to focus on ethical practice for testing professionals. However, according to Spolsky (1995), from the 1910s to the 1960s, social, economic, political, and personal concerns among key language-testing professionals in the United States and the United Kingdom dominated boardroom meetings and decisions. But in the last two decades, ethical concerns have emerged in the research literature. Spolsky (1981) argued that tests should be labeled 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, emphasizing that language testers should be responsible for ethical use of the information they collect. Stansfield (1993) argued that professional standards and a code of practice are ways to bring about ethical behavior among testers. Alderson et al. (1995) reviewed principles and standards but concluded "language testing still lacks any agreed standards by which language tests can be evaluated, compared, or selected" (p. 259). Corson (1997), broadly addressing applied linguists, made a case for the development of a framework of ethical principles by considering three principles: the principal of equal treatment, the principle of respect for persons, and the principle of benefit maximization.

In the last few years, momentum has gathered through publications such as the special issue of Language Testing guest-edited by Alan Davies (1997) and conferences such as the Language Assessment Ethics Conference (see Kunnan, 2002). Narrowing the discussion to applied ethics, Hamp-Lyons' (1997) asked what the principle was against which the ethicality of a test was to be judged. To offer guidance in this matter, the International Language Testing Association (ILTA) recently published a report of the Task Force on Testing Standards (1995) and the ILTA's Code of Ethics (2000) that lays out broad guidelines of how professionals should conduct themselves.

As a way of forward, then, the key question for assessment ethics should be: How should ethical test development and test use be promoted? The answer is that there are indirect and perhaps difficult ways. One way is to develop appropriate group ethics through the development of professional standards (for researchers, practitioners, and students) as recently provided by ILTA's Code of Ethics. This Code could be used in developing, researching, maintaining, and evaluating tests as well as in training graduate students although it may not be universally applicable and adjustments may need to be in certain communities.19

CONCLUSION

This chapter discusses language tests and testing practices by situating them within four main wider contexts. The four contexts discussed here do not preclude the relevance of other contexts that may be useful in other discussions. This chapter also explicitly puts forward the notion that tests are best understood when a wider interdisciplinary perspective is used in debating and evaluating tests and testing practice. As Spolsky (1995) asserts, "In the study of fields like language testing and teaching, scholars need to be ready to draw not just on the obvious theoretical disciplines that underpin applied linguistics, such as the various language sciences and education, but also on fields like economics, political science, and sociology that furnish methods of investigating the context in which language and education exist" (p. 3).
I would like to conclude with McNamara's (2000) beautiful theater analogy that perfectly captures my own interest in the wider context and the spirit of this chapter: At a moment of dramatic intensity in the theater, the glare of a single spotlight can isolate an individual actor from his or her surroundings. The spotlight focuses the spectators' attention on the psychological state of the character being portrayed. Temporarily at least, the surroundings, including other actors present, are rendered invisible for the audience. Until fairly recently, thinking about language assessment was like this. It focused exclusively on the skills and abilities of the individual being assessed... But what does this spotlight of this individualizing perspective exclude? What lies behind, around? Imagine the spotlight going off to be replaced by normal stage lighting: The other actors on the stage are revealed. Now imagine the performance continuing, but the house lights coming up, so that the audience is revealed. Imagine finally the side curtains behind pulled back and the stage removed to expose all the personnel working behind the scenes. The individual performer is now exposed as forming part of a larger collective activity... (p. 67).

NOTES

1. As I am making no distinction in this chapter between tests and assessments, I am using tests interchangeably with assessments and testing interchangeably with assessment.


3. In this view, 'a fair test' is not necessarily one in which a test-taker is awarded a better score.

4. In Fig. 43.1, the four main contexts are presented in neatly overlapping circles. In reality, it is likely that one or two contexts may overlap fully or only be tangentially involved. Much depends on the how and why a test is commissioned, developed, administered, scored, reported, researched, and used by the community in which the test operates. Moreover, the main contexts need not always be configured in the manner shown. In particular communities, depending on the local situation, the main contexts of interests could be social and political or economic and technological, or legal and political, etc.

5. Davidson and Lynch, (2002) in their book Testcraft use the concept of mandate, which they define as a constellation of forces that shape the design of a test: theory, curriculum, politics, money, answerability (to higher powers), etc.

6. Most of the examples in this chapter are from the United States this is where the author works but this does not mean that the "Wider Context" perspective promoted in the chapter is restricted to the United States. This framework could be used for evaluation of tests and testing practices in any community.

7. Most commentators read private schools to mean faith-based (meaning, religious and Christian) schools which is very much in line with the Republican President Bush's agenda.

8. It must be noted that California teachers are not directly involved in any of the standardized tests used in the state. The CAHSEE, the CELDT, or the California Achievement Test 6 are all produced by testing companies thus increasing the divide between teaching and testing and teachers and tests.


10. This website provides a list of about 200 unconnected English sentences for the written English test, much like a set of sentences in an old-fashioned dictation test; test-takers are cautioned that other sentences may be on the test: http://www.ins.usdoj.gov/graphics/services/natz/natzsamp.htm retrieved on 1/25/2003.

11. At the time of writing, the College Board, which owns the SAT, is planning to revise the test. The new test will have Critical Reading (an increase in the reading part and the elimination of the word analogies section), an Essay Writing section (a new addition), and the Mathematics section.


13. The FairTest: The National Center for Fair and Open Testing at the following website has many arguments on these lines: www.fairtest.org.

14. A caveat regarding this type of bias is that outcome differences do not automatically mean that the test is biased but the burden is on the test developer to demonstrate that biased test items are not causing the differential performance.

15. It was reported in U.S. press that the Computer-based TOEFL, which had been launched in sub-Saharan Africa in 2000, soon had to be replaced with the paper-and-pencil version due to problems of geographical access after many university and college professors in departments of African studies in the United States (known as Africanists) complained to ETS about this.
16. This is a uniquely American legal provision due to legal (for example, slavery) and illegal discrimination of several groups of people (Native Americans, African Americans, Hispanic Americans, and Asian Americans) practiced for centuries in the country.

17. Similar laws are on the books in the United Kingdom the European Union, and India.

18. See Bersoff (1981) and Fulcher & Bamford (1988) for legal challenges that are possible in the United States and United Kingdom. Also, see Kunnan (2000) for a list of court cases related to testing in the United States and Lippi-Green (1997) for court cases related to employment-related language use and discrimination in the United States.


REFERENCES


Handbook of Research in Second Language Teaching and Learning

Edited by

Eli Hinkel

Seattle University