

## Ethics in Language Testing\*

Hossein Farhady  
Iran University of Science and Technology

### Introduction

In the last few decades, testing in general, and language testing in particular, has witnessed shifts of focus in theory as well as in practice. The short history of language testing has witnessed different perspectives that have appeared, grown, popularized, taken control over the field, and eventually faded away. Formerly, testing and measurement followed what can be called the instructional perspective. That is, tests were normally prepared and used by teachers only. No scientific characteristic such as reliability or validity was required for the tests used by the teachers as testers. Thus, testing was performed on the basis of the tastes, knowledge, or in short, on the basis of idiosyncratic preferences of the teachers. With the advancements in the application of scientific findings to instruction, testing moved into what can be called the psychometric perspective. Principles of psychometrics were firmly applied to instructional tests with the belief that educational outcome should be measured by valid and reliable instruments. In the present century, psychometric dynasty have ruled the educational measurement.

It should be mentioned, at the outset, that there is a major difference between tests of subject matter areas and those of language. In most educational disciplines dealing with subject matter areas, application of psychometric principles would probably be sufficient to make the tests acceptable because the subject to be tested is relatively clear and unambiguous. A major advantage of tests in subject matter areas is that in such tests, the medium of testing, i.e., language, is not of major concern to the testers. Nor is it problematic to the test takers because such tests are often designed to measure the degree of the students' knowledge in a particular area through the test takers' native language. Nevertheless, this does not imply that tests in such fields are flawless. As Bachman (1990) states, the concepts of imperfection, incompleteness, and subjectivity are always intermingled with the nature of testing and measurement. In the field of language testing, in addition to the above-mentioned problems, there is one more complexity. That is, the testers have to measure the degree of the testees' command of the language through the medium of language. Therefore, what is being tested is confounded in the medium of testing itself. In other words, language tests are intended to measure the construct of language ability that has not been clearly defined on the one hand, and to measure and possibly remove the overlap between the language of the test and the language to be tested on the other.

Within the psychometric perspective of language testing, then, two major obstacles have persisted. First, an operational definition of the phenomenon to be tested has not been available. This lack of concise operational definition is due, among other reasons, to the complexity, the multidisciplinary nature, and the abstractness of the construct of language ability. Once defined in pure linguistic terms, language ability is now an applied linguistic phenomenon influenced by the advances in psychology,

sociology, anthropology, and other related fields. Consequently, the formerly accepted linguistic definition of language ability is no longer satisfactory because defining language ability in purely linguistic terms does not seem comprehensive enough to account for its use, change, acquisition, and interpretation. These aspects, which are not exclusively related to linguistics, constitute the interdisciplinary nature of language as a medium of communication. The use of language, for example, relates to sociolinguistics, its changing nature to linguistics, its acquisition to psycholinguistics, and the interpretation aspect to discourse and pragmatics. That is why a comprehensive treatment of language through one single dimension is neither easy nor acceptable. In addition, concepts and constructs in the above-mentioned areas, as branches of applied linguistics, are not concrete or objective by themselves. This subjectivity adds more complexity to the definition of language itself as well as to that of language ability. When the ability to be measured cannot be, at least, easily defined, the whole idea of validity, which is the heart of the educational measurement, would be jeopardized.

The second shortcoming in the psychometric perspective is rooted in the problems and prospects of the measurement field itself. From among theoretical approaches to measurement in educational psychology, Classical Test Theory dominated the test statistics for a long time. The scope of this paper does not allow for a discussion of the deficiencies of the Classical Test Theory; nor does it seem necessary here because it has been treated quite comprehensively in the literature (Henning, 1987; Bachman 1990). However, although the emergence of Generalizability Theory and Item Response Theory to account for psychometric problems of measurement have considerably advanced our understanding of the test score treatment, they have not cured all the ills either (Shavelson & Webb, 1991; Hambleton, et al. 1991; McNamara, 1990, 1991, 1996). This implies that regarding the measurement of language ability, some imprecision exist which could jeopardize the reliability, the brain in contrast to the heart, of the tests. Thus, from a live creature like a test of which the heart and the brain do not or cannot function properly, a perfect performance should not be expected.

It should be noted, however, that no matter how imprecise psychometric devices of language measurement might be, and no matter how applied linguistics foundations of language tests might be, language educators have to prepare and administer the tests. Furthermore, language testers are bestowed with the power of making decisions on the educational, social, occupational, and sometimes on the personal lives of the test takers. That is why, tests in general, language tests no exception, have turned into extremely powerful weapons in the hands of not only teachers, but also administrators, economists, politicians, and some other authorities in the society.

When there is power, there develop all sorts of principles for the appropriate implementation of the power. In language testing, “appropriate implementation” is often referred to as fair application of test results to the decisions made on the lives of the test takers. Fortunately, in the last few decades, an appreciable attention is directed towards a fair implementation of test results under the topic of **ethics** in language testing (Lynch, 1996; and Shohamy, 1997). However, as with many other concepts in the field of language testing, there are disagreements among the scholars on the treatment of the term ethics from different perspectives. The issue is touched

upon here from a sociopolitical perspective. More specifically, this paper addresses the following issues. First, the uses and misuses of language tests that may lead to unfair decisions, which in turn, may raise the question of ethics in language testing, will be presented. Second, the concept of ethical and unethical applications of test results within different sociopolitical contexts will be explained, arguing for the idea that, in some cases, an unethical perspective may be not only inevitable, but also necessary. Finally, some suggestions and guidelines will be offered for fair implementations of language tests.

## Uses and Misuses of Language Tests

Extensive and strict application of psychometric principles led educational testing into a discipline where numbers played a determining role. From the first stage of developing a test to the last stage of interpretation of test scores and evaluation, numerical information controlled the whole process of testing and measurement. Items were prepared, revised, and evaluated with reference to numerical values such as item facility, item discrimination, and choice distribution. Tests were judged almost entirely on the basis of empirical pieces of information such as reliability and validity indexes. Decisions were made on the basis of the numerical values obtained from the tests. This number oriented attitude toward language testing became so strong that sometimes it overshadowed the quality of the tests in terms of content relevance and context appropriacy.

On the other hand, the diversity of tests, developed under different conditions and used for different purposes, led to the multiplicity of and the discrepancies in the interpretations of the numbers obtained from the tests. Thus, test organizations tried to justify their interpretations by resorting to psychometric qualities of the tests. For instance, it was assumed that the higher the reliability and validity indexes, the better the test. Therefore, test developers tried their best to improve such numerical values even to the cost of quality. However, when the interpretations of the scores expanded too much, they tended to confuse the test users as well as the decision makers. As a result, the field was forced to come up with a uniform interpretation of the numbers because the reported numbers had to communicate similar meanings to different people independent of the effect of contextual and situational variables. Apparently, the idea of uniformity in its strong sense led to the concept of standardization. Thus, an era of standard tests started, developed, and extended control over the entire educational systems in the world.

The era of standardization mandated a clear definition of the concept of the "standard" itself. That is, both test administrators and test users were interested in an agreed upon definition of the expression of "a standard test". At first, standardization was used as an umbrella term for a variety of tests even with unclear and unspecified characteristics without any serious objections from the field. With growing dissatisfaction about the concept of standardization, test developers began to identify the extent of standardization for tests, especially for professionally marketed, and I don't mean professionally catered, tests. Eventually, some features were identified as necessary requirements of standard tests. These features included uniform procedures for test planning, item selection, scale development, norming, instructions to the testees and proctors, timing, scoring and interpretation of the scores. Each and every

one of these issues has been the focus of much discussion in the literature; however, a brief treatment seems warranted here.

Standardization in planning a test required a comprehensive and, at the same time, detailed review of the materials which would constitute the corpus for the content of the tests. Of course, in the case of attainment tests such as achievement, it might not seem so difficult to exhaust the content of the instructional materials. The problem for different testing organizations, however, was to treat the materials in a uniform fashion to come up with similar specifications for the content of the test as well as the same weights to be assigned to different parts of the materials. In general cases such as proficiency tests, the problem is undoubtedly intensified, because in addition to the difficulties associated with achievement type tests, the task of specifying the content of the corpus, the form in which the test has to be developed, and the arrangement of the items would add to the complexity of the procedure. Therefore, in most cases scholars have been forced to utilize subjective and sometimes arbitrary specifications of the content of proficiency tests. Although such a problem is not exclusive to language tests, it is manifested itself more forcefully in language tests than in the tests prepared for other academic fields because the domain of the content of language tests is virtually unlimited.

Standardization in item selection required homogeneous procedures in order to include a quite representative sample of the content of the materials in the test. Representativeness is not a completely objective phenomenon by itself. From the same corpus of data, albeit prepared carefully, different tests can be developed all of which can be called representatives of the same corpus. However, depending on the theoretical perspective of the test developers within a particular educational context, different selections, even randomly done, may lead to dissimilar tests. Furthermore, the tests may turn out to be even more dissimilar if the items are prepared with certain preconceptions regarding the educational background of the test takers in mind. For instance, in the case of centralized education systems, a representative sample of items for one region in a country may not be representative for another region due to different qualities of education in different parts of a country. The case would be more problematic in decentralized educational systems. Of course, the inequality of the content of the tests would lead to bias in the testees' performance and eventually to unfair decisions on the test takers.

Developing a measurement scale is another issue in the process of standardization. At some stages of standardization, certain assumptions are made some of which may not be justified. For instance, all items are assigned equal credits assuming that they have similar instructional value or they demand equal cognitive load on the part of test takers. Therefore, in spite of the fact that items often have differing instructional values and impose differing mental loads on the testees, they are all given the same credit. This has led to a commonly agreed upon scaling of all standardized tests. That is, almost all abilities are supposedly measured on interval scales. It should be pointed out, however, that interval scaling suffers, in its very nature, from inequality of the intervals.

In interval scaling, it is assumed that the distance between the intervals is equal, i.e., the difference between the scores of 10 and 11 on a test is assumed to be the same as

the difference between the scores of 90 and 91. This assumption is far from reality in the context of teaching, learning, and testing. The amount and probably the value of the knowledge which may push a testee's score from 10 to 11, is far less significant than the amount and the value of the knowledge which would push the same testee's score from 90 to 91. That is, scores on psychological tests, all of which are based on interval scale, rarely provide absolute, ratio scale on the measurement of the attribute. Therefore, it may not be meaningful to claim, how much of an attribute exists, but how much of a particular ability an individual has in relation to or in comparison with a norm that provides standards for interpreting test scores. Despite such complexities, however, the interval scale has continued to be the predominant scale for almost all tests as if it were flawless and quite appropriate for measuring language abilities.

When the test is planned, items are selected, and the scale of measurement is determined, one has to decide on the norming of the test. Normalizing has a long history in the educational context. What is worth mentioning here is the importance of the range and the scope of standardization. Local norms, national norms, and international norms demand different degrees of vigor and rigor in their preparation and application. Of course, the narrower the norm in terms of its coverage, the easier the norming will be. For large-scale normings, the sample should be quite representative of the population. Otherwise, non-representative samples would most likely lead to inappropriate norming and imprecise decisions on the basis of test scores.

Although standardization through the above-mentioned stages of test development may seem obvious for testing authorities, not all the stages are followed to the satisfaction in some standardization processes. When such well known factors have not been satisfactorily accounted for, other less obvious ones may not even be taken into consideration. For instance, it is assumed that instructions to the proctors, testees, and administrators are interpreted exactly and accurately by all people involved in testing processes. Almost all contextual, situational, and environmental varieties that might influence the examinees' performance and thus introduce irrelevant score variance, are most often ignored. Although the purpose of standardization is to eliminate as many extraneous factors as possible from affecting test performance, most of these factors along with others such as physical surroundings in which the test is done, the health, attitude, and emotional status of the examinees, time of the testing, to name a few, cannot be empirically accounted for. Nor could they be standardized. In most cases, these factors creep into test development and administration processes and contaminate the information that is supposedly related only to the trait being tested.

In addition to the above-mentioned problems, which were mainly related to the process of test standardization, other disadvantages have often been attributed to standard tests. The most significant of all, is the correspondence between the content of the standard tests and that of what the learners have actually been taught. In most cases, students focus on learning the materials that are most likely to appear in the test. Apart from the fact that this is the student's legitimate right to make best out of his endeavor in preparing for the test, it may lead to what Livingston, et al. (1989) called "test driven curricula". They claim that when the results of the tests are used to compare individual learners, teachers, school systems, and even school districts, it

might be assumed that test driven curricula would be the most logically natural outcome (Koehler, 1978; Rice & Higgins, 1982; Wood, 1982; Tindal, 1983). Most of the scholars, however, claim that there is a mismatch between the content of the standard tests and that of what exactly goes on in the classroom.

Moreover, some testing authorities claim that nation-wide tests are designed to measure students' achievement in schools. The purpose of such tests is to improve the quality of education. This implies that the testers control and almost dictate what will be tested and consequently what will be learned. Teachers and educators, on the other hand claim that testers view tests as synonymous with curriculum and to learning (Shepard, 1991). Nation-wide tests are often based on the materials to which not all schools are equally exposed. They claim that such tests are inefficient because there is no correspondence between what teachers are supposed to teach, what they actually teach, and how their teaching outcomes are evaluated.

Finally, some scholars have demonstrated that factors such as sex, socio-economic status, test form, students' majors, and exposure to language, significantly influence test performance. For instance, Shohamy (1992) examined the effect of the introduction of a new EFL oral test as part of the nation-wide high school matriculation exam on the level of English proficiency. She found that introducing a new test led to the teaching of "test language", i.e., only those language tasks and skills that were likely to appear on the tests were taught. Furthermore, teachers focused on the procedures that facilitated achieving the goals of the test rather than on the real learning of the materials. Further, it was claimed that performance on a standard test and meeting the standard set by the test does not necessarily guarantee the existence of the abilities which the testees are required to perform. Thus, standard tests served neither the purpose of achievement tests nor the purpose of proficiency tests.

In spite of the existence and persistence of the above-mentioned problems, most scholars claim that standard tests are efficient devices to measure educational outcomes, at least more efficient than the traditional assessment procedures. Of course, this is a tradition in educational measurement and specialists in this area believe that mass assessment is a procedure to investigate the failure or the decline of educational programs and to offer remedial steps. They also believe that standard tests are used to provide comparable data for the purposes of offering guidelines for students' future academic career, and even to investigate the efficiency of instruction (Wiggins, 1989; Robinson & Craver; 1989). Others go further to claim that standardized tests render viable, inexpensive, reliable, and valid indicators of student learning. They also claim that data from standardized tests are readily available, cheap, and abundant. Most statistical properties are provided by the test developers which make test users relieved from the complexities of data analysis. They further believe that standardized tests make it possible to generalize and to draw conclusions about the data and their implications (Sanders & Horn, 1995).

Although there are advantages and disadvantages regarding the utilization of standard tests, they still have their position as major sources of information upon which decisions are made. When there is a decision to be made, there comes the power on the side of the decision makers. That is why tests (standardized or teacher made) with

local, national, or international norms have become a means of power in the hands of people who may not be well informed about the process of testing and measurement. Since uses and misuses of power can lead to serious consequences, a brief description of the concept of power in language testing seems necessary.

### Power of the Tests

There is no doubt that testing is an indispensable and, at the same time, an important part of education. Therefore, those who are involved in education (teachers, learners, administrators, teacher trainers, and decision makers) assume responsibility toward their profession. When there is responsibility, there should also be power to enforce the fulfillment of the responsibility. Thus, tests are and should, I would assume, by their very nature, be powerful. That is why Spolsky (1997) claims that the concept of power has been inherently associated with tests and examinations ever since their invention.

Part of the power of the test is due to the fact that within the areas of human sciences and particularly applied linguistics, testing is one of the few fields that approximate empirical sciences. Testing is scientific and allows experimental procedures through which empirical data can be obtained and statistical techniques applied (Shohamy, 1991). That is, the collected data would be interpreted as objective and quite true. The “objective” and “empirical” information obtained through tests is then used as a reliable device for making decisions. She also argues that tests are powerful because (a) they produce scores that are possessed by the testers not the test takers, and documentation of these scores places the individual in an area of surveillance; (b) they are described by decision makers as useful educational means for the advancement and improvement of education. That is, through the tests, educational authorities exert power over the educational systems to make their intended modifications; and (c) testers assume that test scores are obtained through objective measures and that these scores provide true pieces of information on test takers’ ability (Shohamy, 1997). Along the same lines, McIntyre (1984) states that since the aim of decision makers is to adjust means to ends in the most economical and efficient way, they will use scientific information as if it were universally true

In addition, tests are powerful because so many people are influenced by test scores. These people, or to follow Rea Dickins (1997) stakeholders, include language testers and teachers, parents, administrators, teacher educators, governments, sponsors and funding organizations, public, and the list is not exhaustive. Each and every group has a share of power in testing process. The long list of stakeholders can be divided into educational, social, and political groups. Each group has a different type of interest in and intention for utilizing tests as a source of power because each group is concerned with a number of internal and external factors.

People in the educational context include students, teachers, and teacher trainers whose careers are often evaluated by the results of the tests. They are the ones who might be later victims of the tests being administered. Therefore, they show more enthusiasm for the test quality and are anxious about the consequences of the tests. More specifically, teachers do know what they have taught, what portions of the materials they have concentrated on, and what subjects deserve to be included in the

test. Therefore, they are curious to observe the correspondence between what they have taught to the students and what the students are tested on.

The discrepancies between what is taught and what is tested would lead to changes in instruction, in focus on the materials, and eventually in the quality of education. Thus, whether intended or not, tests administered on a national scale bring about certain modifications in the quality and the quantity of education. These changes are desirable if they are in the direction of improvement though, in some cases, they are not purely educational. Further, tests are sometimes used by the authoritative agencies to frighten the educators in order to stimulate the system. It usually occurs when there is a failure in learning. An example is the proposals to introduce national tests in the USA because of the low scores that American children obtain on international tests (Madaus, 1991). Spolsky (1997) also points out that test results are used as tools for enforcing educational goals, especially in situations where actual education has failed.

People in the social context are greater in number than those in the educational context. From individual and family characteristics, to community structures, to national and international macrostructures are all influential factors regarding the implementation of power by the test users. Test takers, as members of different families, react differently to test results because they have different infrastructures. Some individuals do not have much sensitivity toward the outcome of the test because their family structure makes no imposition on the young members of the family to try hard enough to pass a test. Even some families are not very much concerned about what the consequences of the test scores might be on the lives of their children. On the other hand, there are individuals whose families are quite dogmatic about the test scores. Such families create the impression on their children that a failure on a test may cost their whole future. Therefore, these families would exercise all sorts of restrictions and limitations on the personal lives of the youngsters in order for them to succeed in the tests. And still a great majority fall in between the two extreme positions. The various attitudes toward test results can be attributed to many factors including the educational background of the parents, the socio-economic status of the family, the number of children in the family, and even the residential area of the family.

The community structure is also important in shaping the attitudes of its members toward how powerful tests can be. In some countries, there is an incredible competition among the applicants in certain cases. For instance, one and half a million high school graduates take the university entrance examination in Iran where only about one tenth of the applicants can have their ways to higher education. For some of them, it takes one or two years of round the clock preparation for the test which usually costs them more than the total annual income of their families. In some communities, even the personal and social values of individuals are judged by their success or failure in this particular exam. Thus, there is an immense degree of pressure on the individuals and families in the community regarding the outcome of the tests.

National and international factors are also important in some countries. The number of students in the higher education institutes, i.e., the quantity, and the number of students winning some international academic awards, i.e., quality, leave a great

pressure on the educational authorities. This pressure is downloaded to teachers to prepare students for such competitions. In some cases, these students are selected through different phases of local and national testing procedures. Those who are selected, go under strict and special training to succeed in international competitions. Although success of such specially trained people, who do not represent the population, is in no way an indication of academic standards of the society, it is recorded as national academic achievement. On the other hand, indexes referring to the number of academic institutions, students, teachers, and the international awards won by a few intelligent and specially trained people often determine the academic rating of a nation in the world.

People in the political context probably enjoy the highest share of power regarding the outcome of the tests. Those who make educational policies at the national levels can make decisions that might turn the test scores into a source of fear for the individuals. The exercise of power and control through tests can be observed in a variety of settings and contexts. Tests are used by policy makers as tools to manipulate the educational system, to control curricula and to impose the introduction of new textbooks and new teaching methods.

For instance, Hawthorne (1994) claims that Australia can be characterized by the use of language testing for political purposes - frequently in the context governed by macropolitical pressures (Hawthorne, 1994, 1995; McNamara, 1990, 1996). English language ability is used as a criterion to control skilled immigration intakes. When the country was suffering from economic recession and the government wanted to cut the number of immigrants, high pass levels were required. With the economic recovery and a cautious rise in the migration program, language requirements were significantly eased (Hawthorne, 1994). Another case is the standard set by universities in western countries for the admission of nonnative students. For example, ETS has set the standard of 500 on TOEFL as an indication of one's ability to pursue one's education in a university where English is the medium of instruction. When Iran was exercising some political problems with these countries, some of the universities raised the standard from 500 to 550 just to deny admission to Iranian applicants.

Educational and sociopolitical contexts mentioned above normally inject power into the tests that is somewhat worrying because there is always a possibility of the misuse of power. It does not really make any difference who misuses the power because the result will harm an individual or a group of people. Of course, all tests including language tests are constructed by a specialist or a group of specialists in an academic atmosphere and with good will. They hope that their tests will be used fairly and justly. However, the uses and misuses of the tests as powerful tools for decision-making are not under the control of the test developers.

A potentially reasonable way to avoid power misuse in testing is to inform those in power through giving them information on the nature and consequences of the tests. Then, it would be possible to control the misuse of power in a logical manner. For example, some stakeholders are important. The more important ones make the decisions and take actions while the less important ones are those affected by the decisions. Therefore, the orientation should be directed toward the people in the

power position in order to help them make fair decisions. Those who are affected by test scores also need orientation to understand how they have been judged and what the test scores could mean to them. In spite of the fact that language testers have repeatedly and publicly announced the importance of decision making, many decisions are still made which cannot be justified on the grounds of the ability being tested. As Shohamy (1997) states, the difference between description and judgment should be taken into account in decision making; while tests give descriptive information, decision makers often use the results for judgment, i.e., punishment or reward.

Since the use of tests for the purpose of power and control has become a widespread phenomenon in many countries, questions have been raised not only on whether using tests in such a way can advance and improve learning in a meaningful way, but also the ethicality of using tests for these purposes. That is why ethics in language testing has become one of the major concerns of language testers in recent years.

### Ethics in Language Testing

Problems in language testing are not limited to only the vagueness of the trait being measured or the imprecision in the measurement of the trait. The ultimate goal of administering a test, whether it measures a well-defined construct or not, and whether it measures the construct with precision or not, is to enable test authorities to make fair decisions on the lives of stakeholders. Bachman (1991) groups the types of decisions to be made into two major categories. The first category refers to the decisions made about test takers' language abilities as well as their ability to use language in contexts outside the test itself. The second category refers to a set of different decisions. They include the decisions made on selecting the testees for a particular occupation or academic career, on diagnosing the students problem areas in language, on placing test takers in appropriate channels of instructional programs, on monitoring student progress, on assigning them class grades, on granting them certificates of many kinds, and on giving them occupational and employment opportunities.

In spite of the multitude of decisions to be made, language testers are often not directly involved in decision-making processes. However, they are aware of the potential problems with test scores due to unclear definition of the trait as well as to the imprecise measurement procedures. They are also aware that no test score is, in any sense, an absolute indication of any ability it may claim to be. Therefore, had language testers been in charge of decision-making, they would have made decisions cautiously. Nevertheless, decision makers are usually, administrators, bureaucrats, and politicians who are not well aware of the problems involved in testing. They often assume that test scores are true indications of abilities and make decisions as if there were no faults in testing procedures. That is why language testers have always complained about the total authoritarian attitude of the decision makers. Language testers claim that decisions should be made with taking the potential flaws of test scores into consideration. Language testers also claim that the decisions should be fair, not harming, and just regarding the test takers. In recent years, these issues in language testing have been receiving an increasing attention under the cover term of ethics.

No one would deny the fact that ethical considerations are important, not only in language testing, but also in all academic activities. Nor would anyone disagree that language testers should assume responsibility toward social and individual aspects of the test takers' lives. What is difficult to operationalize, however, is the definition, extent, and limits of ethics. Punch (1994) claims that ethical issues include consent, deception, privacy, confidentiality, and equal opportunity to learn. Lynch (1997) elaborates on these concepts quite succinctly. However, neither one clarifies the context of discussion, i.e., whether language testing is for the purposes of making decisions, or language testing is for the purposes of conducting research.

Regarding ethical issues, then, some confusion exist. Concepts such as fairness, bias, morality, and the context in which these concepts have been used, are not at all clearly identified in the field. Fairness is discussed in terms of bias, and bias in terms of ethics, and both are considered immoral. Even Lynch goes too far to claim that Davies' definition of a test as a discriminating instrument is immoral to start with. Davies (1997a) defines a test quite professionally because in the context of assessment if there is no discrimination, there is neither fairness nor morality. For instance, if calling a test taker "a low achiever" might hurt his/her feelings morally, not telling a test taker a "high achiever" might hurt his/her feelings even more because the knowledgeable test taker deserves to be informed of his high achievements.

In relation to consent, Lynch states, "this may not be a matter of much concern in the context of language testing, except when language tests are used for research purposes" (p. 317). However, the effect of other parameters of ethicality such as deception, privacy, and confidentiality seems to be, at best, unclear. Almost all test takers are well aware that they are supposed to take a test, and in most cases, they are informed of the type, purpose, and consequences of the test in advance.

Furthermore, the importance of the consequences of the test results, what Messick (1989) calls consequential validity, seems to have been overemphasized in recent years. Davies (1997b) states, quite considerably, that the apparent open-ended offer of consequential validity goes too far. He maintains that it is not possible for a tester as a member of profession to take account of all possible social consequences (p. 335). It can be argued that in some sociopolitical contexts, a tester is not even allowed to assume responsibility towards the consequences of the decisions made on the test scores. In such cases, language testers are assigned to develop such and such tests and deliver the tests to certain government agencies. What kind of use the agencies would make from the test scores is often a matter of sociopolitical concerns rather than an academic endeavor. Thus, ethical considerations should be discussed in different contexts each with different parameters because what may be ethical in one context may not be so in another context. Some of these contexts are educational contexts (research vs. decision-making), sociopolitical contexts (public vs. government), and moral contexts (fairness vs. bias). Each is briefly discussed below.

### Educational Context (Research vs. Decision Making)

*Educational context* refers to what Bachman (1990), and Bachman & Palmer (1996) call *test use*, or the purpose for which a given test is administered. The idea of ethics

and ethical considerations has long been established as an important issue in the context of educational research. Right of being anonymous, right of privacy, and right of confidentiality, are agreed upon codes of conduct in educational research. Although these ethical considerations are necessary in conducting research, they are not essential in the context of language testing for two reasons.

First, in doing research, the main purpose is to collect information in order to uncover the mysteries and solve some of the problems of human life. There is absolutely no force on the subjects to participate in research because no decision is to be made on them personally. Second, since in the research context, people who serve as respondents or participants, are not the target for decisions, the researcher is mostly interested in the nature of the information rather than its exact source. Therefore, the identity of the respondents is not as important as the quality of the data they provide. Only in special cases of outliers would the researchers be curious about the personal identity of the subjects for the purpose of follow up studies. In other words, for research purposes, one can enslave a number of subjects and collect data on various aspects of a variable. Depending on the extent of internal and external validity of the research project, a certain degree of generalizability can be achieved.

In the context of educational testing, however, which often ends up with some sort of decision-making, however, testers should be concerned with the realities of educational contexts rather than with theoretical rhetoric of testing. No matter how psychometrically sound, theoretically strong, and ethically desirable a test may be, practical considerations may limit its use. That is, from the four desirable characteristics of a test, i. e. reliability, validity, ethicality, and practicality, the last one may override the others depending on the limitations and restrictions of the educational systems. The following examples may help clarify the point.

No one would disagree that listening comprehension is an important component of language ability, especially in an academic setting. Nor would any board of education disagree with incorporating a listening comprehension component into the language test batteries. To administer a listening comprehension test, however, the least a practitioner needs is a tape recorder, a native or a near native speaker of the language to prepare the tape, and more importantly, a dependable power. When such rudimentary facilities are not available in a particular testing context, it would not be fair to include a listening comprehension component in the test.

As another case, although oral interview is known to be a valid and reliable measure of speaking ability, most standardized language tests deliberately ignore measuring this ability because it has practicality constraints. In most cases, regardless of these limitations, however, the educator or the language tester has to make decisions on a large number of applicants. This is exactly where the social responsibility of the language tester plays a crucial role in offering a viable alternative. If indirect measures can be utilized to compensate for these constraints, the outsider critic would consider the test as unfair and the decisions unethical, while the whole idea was to improve the quality of the test scores in order to make fairer decisions.

Thus, the concept of ethics is and should be important in both testing for research purposes and testing for decision-making. The point, however, is that in the context of

education, some factors may influence the interpretation of ethicality beyond its commonly assumed limits. More importantly, it should be determined whether ethical considerations in testing are similar to those in research or testing should find its ethical considerations and standards in its own context.

### **Sociopolitical Context (Public vs. Government)**

Within the social contexts of language testing, the degree of public awareness about the quality of the tests is an important factor. Public interest in the quality of the tests, however, depends very much on the success or the failure of the test takers. That is, no matter how ill-formed a test might be, as long as the test taker obtains a passing score, no one would complain or even be concerned about the test. Failure on the test, however, would be often blamed on the quality and unfairness of the test rather than the test takers ability. This is particularly true in the societies where the culture of testing is not appropriately cultivated. In such societies, there is a wide gap between the public perception of what a language test is and what the language test does.

Political context is even more important because it is related to the government agencies. Governments are often the owners of the educational institutions, and thus act as super controllers of the whole systems. For instance, certain laws are suggested and approved by the parliament to give certain stakeholders an advantage on the test over the others. The reason for such a decision may not need to be explained here. It would suffice to state that no matter how legitimate such decisions might be, they are not academically justified. Of course, within the conceptual world of ethics, such a procedure is immoral, unethical, biased, unfair, and a legalized illegality. Nevertheless, it is a fact that political preferences mandate certain actions that are quite justified on the basis of the conditions of a country. Thus, what might be considered quite unethical somewhere is perfectly ethical elsewhere. This implies that ethics is simply a value code rather than an objectively defined term.

### **Moral Contexts (Bias vs. Fairness)**

Morality refers to whether a test is biased for or against a particular group of test takers that might lead to unfair decisions. It should be mentioned that the terms bias and fairness are closely related but quite distinct at the same time. Bias is a statistical characteristic of the test score, or of the predictions based upon those scores. Bias is said to exist when a test involves systematic sources of error in measurement or prediction. The existence of bias can be defined empirically and determined statistically. By examining the data, one can determine the extent to which a test provides biased measures or biased predictions. Fairness, on the other hand, refers to a value judgment regarding decisions or actions taken as the result of test scores. It involves a comparison between the decisions that were made and the decisions that should have been made. A test is most likely to be attacked as unfair when (a) it leads to adverse decisions for some groups in the population, (b) it is the sole basis for decisions, and (c) the consequences of doing poorly on the test are harsh (Seymour, 1988).

Although fairness cannot be determined by statistical methods, it can be remedied in different ways. Spolsky (1995) reviews the literature and states that in old times, one way to mitigate the hardship of those who having just approached the gates of Paradise, are excluded. Although in all probability, it would not a whit worse than

those who are just included, by giving a second examination to the best of the unsuccessful and the worst of the successful, and ranking them on the average of the two results. Thus, one way to alleviate the consequences of unfairness is multiple-assessment through which many relevant factors can be taken into account. Another way is to employ multiple stage decision models rather than making irreversible decisions about everyone at the point of testing (Cronbach & Gleser, 1965).

### Conclusions and Suggestions

In order to put the sociopolitical aspects of ethics in language testing in a proper context, a three-way conversation between the test taker, the test developer, and the decision maker is presented.

The test taker states: "I am a high school graduate. I have studied English for 6 years. The kind of English I have learned is planned with the materials developed by the central board of education. I usually received the textbooks a couple of months after the academic year had begun. In some cases, I did not have a teacher from the beginning of the course. When I had the textbooks, I did not have the teacher, and when I had the teacher, I did not have the textbook. There is no private language institute in my hometown so that I could improve my English. Nor did I have the facilities to attend even if there were the opportunities. I have had no access to extra educational materials, books, magazines, technomedia, or any other facilities. I have a friend, however, who is also a high school graduate, whose case is quite different. He attended private institutes ever since he was in junior high school. He studies in a private school with highly qualified teachers with excellent materials in English. He had also access to many TV programs broadcast in English. Although I may be more motivated to learn English than my friend might have been, I have not had the opportunity to learn the language. With this sort of difference in our educational backgrounds, I don't think it is fair to give both of us the same test because he would naturally outperform me".

The verse seems quite impressive and may raise a lot of enthusiasm from the audience. Everybody would tend to believe that given the same test to such differently educated people might not be fair.

The second person in the conversation is the test developer who claims: "I am in charge of test development. My responsibility is to prepare a test on the basis of the specifications given by the officials. I have been assigned to make the test in such a way that it represents the whole materials that were supposed to have been covered during the courses of instruction. I was not provided with information about the regional differences in the quality and the quantity of education. Even if I had the information, I was not permitted to consider such differences in the construction of the test. I don't even teach anymore, and I don't have any idea of the priorities given by different teachers in different regions to different language elements. The test, however, is developed on the assumption (whether true or not, whether fair or not, whether valid or not) that all students in all places in the nation have studied and supposedly learned the materials. Fair or not is not my responsibility."

Such statements make people feel enthusiastic about the test developer because he is not really in fault. He has followed the instruction given to him by the authorities. The

test taker and the test developer seem to be innocently guilty in the testing process that does not seem fair.

The third person in the conversation is the official in charge of the testing process who claims: "Based on the rules and regulations governing the national education, the whole planning is done once and for everybody in the country. The textbooks are the same, the teachers are instructed to cover the materials, and everybody is given similar opportunity to learn. Of course, I am aware of some inequalities at the levels of education across the nation as well as unequal distribution of good teachers and availability of good materials in different places in the country. However, I have to follow the rules of competition and select the most knowledgeable students for the purpose of higher education. It is not very important whether the students have had equal opportunities to learn. What is important for me, as the person in charge of selecting students for the job, giving them admission to universities, or awarding them certain educational grants, is to select the best ones from among the members of the group. If I had to take into account such discrepancies, I had to give different tests of differing degrees of difficulty, or I had to make selections on different qualifications. This would jeopardize national standards and the purpose for which the test is designed.

The audience is left in a dilemma. It is quite logical for the test taker to complain about the opportunities he had had for his educational career. It is also quite natural for the test developer to clean himself of any accusation regarding the unfair testing procedure. Further, it is quite understandable for the officials to make their best effort in selecting the most competent applicants from among others. Then where is the flaw? Who is acting unfairly? Where is the bias? What is unethical? Finally, what is the responsibility of each, in general, and the responsibility of the language tester in particular? Pennycook (1994) answers this last question by stating that, "The responsibility of language testers is clear: we must accept responsibility for all those consequences which we are aware of. Furthermore, there needs to be a set conditions and parameters inside which we are sure of the consequences of our work and we need to develop a conscious agenda to push outward the boundaries of our knowledge of the consequences of language tests and their practices."

To achieve such a goal, language testers should assume three types of responsibility. First, they should attempt to provide a clear definition of the trait being measured, in this case language ability. Clarification of the trait to be measured is the responsibility of the testers in so far as it can be fulfilled by the findings in applied linguistics. Second, language testers should be concerned with how precisely the trait can be measured. The precise measurement of the trait is the responsibility of the language testers in so far as it can be fulfilled by the findings in psychometrics. The third, and probably the most important responsibility of language testers is to exercise care about the decisions made on the basis of scores obtained from the measurement of the trait. The appropriacy of the decisions to be made, which is the social responsibility of the language testers, may be fulfilled in so far as they are involved in the process of decision making.

To fulfill the first responsibility, language testers can benefit from the most advanced developments in applied linguistics. They can utilize the latest theoretical findings about the nature and the structure of language from the field of linguistics. They can

also get advantage from the findings in the area of educational psychology and second language acquisition about the theories of learning in general, and those of language learning/acquisition in particular. By accumulating knowledge from different fields of applied linguistics, language testers feel responsible for formulating the clearest understanding of the trait of language. As for the second responsibility, language testers can utilize the findings of psychometricians to understand the complexities of measurement theories. They assume responsibility in practicing the principles of psychometrics in order to make the measurement of the trait as precise as possible. Regarding the third area of responsibility, i.e., decision making, however, testers cannot do much because decision-making is in the hands of people other than language testers. In spite of the fact that language testers are at the center of the process of language testing, they do not have direct access to the parameters of decision-making. Politicians, educators, bureaucrats, and so many other sectors in the society are in the position to make decisions. Language testers' responsibility in this case is limited to only giving guidelines to the decision makers, of course, if they ever get the opportunity.

The interaction among language testing, applied linguistics, and psychometrics, has helped the field of language testing to mature in recent decades. It has not been more than a three-decade period that language testing has moved from utilizing simple primitive statistical techniques to test data to most sophisticated ones. It may be true that in the early years, concepts such as the mean, the standard deviation, the variance, and correlation coefficient were striking for the people in the field. At present, however, utilizing the most complex statistics such as Generalizability studies, different models of Item Response Theory, structural equations models, and varieties of factor analytic techniques do not surprise any one in the field. All these achievements have been possible through collaborations among language testers, applied linguists, and psychometricians. Such a trend will improve the testing conditions to the benefit of all stakeholders in the future.

A word of caution is in order here to close the paper. Putting the findings of different fields of study together, which is inevitable in interdisciplinary fields, may lead to some disadvantages as well. Some language testers may not have a firm grasp on the complexities of the fields from which they may seek help. This may cause misapplications of the theoretical and practical principles. No one would disagree with the misuses of correlational and factorial analyses in the 70's and 80's, which led to an outbreak of the so-called indivisible nature of language as well as other mental processes that somehow involved language. Therefore, great care should be exercised in the application of the principles of the other fields to language testing in order to avoid misleading conclusions.

This is a fairly revised version of a paper presented at the Summer Institute on The Social Responsibility of Language Testers, Carleton University, Ottawa, Canada (July 1998). It is also printed in *Moddaress*, 3 (11) (1999).

## Bibliography

- Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford: OUP.
- \_\_\_\_\_ (1991). What does language testing have to offer? *TESOL Quarterly*, 26 (4).
- Bachman L. & Palmer, B. (1996). *Language tests in practice*. Oxford: OUP.
- Cronbach, L. J. & Gleser, G. (1965). *Psychological tests and personnel decisions* (2<sup>nd</sup> ed.). Urbana, IL: University of Illinois Press.
- Davies, A. (1991). *Principles of language testing*. Oxford: OUP.
- \_\_\_\_\_ (1997a). Introduction: The limits of ethics in language testing. *Language Testing*, 14 (3).
- \_\_\_\_\_ (1997b). Demands of being professional in language testing. *Language Testing*, 14 (3), 328-339.
- Hambleton, R. K., Swaminathan, H. & Rogers, H. G. (1991). *Fundamentals of item response theory*. Sage, Newbury Park, CA.
- Hawthorn, L. (1994). The politicisation of English: The evolution of language testing. *People and Place*, (2). Melbourne: Australian Forum for population studies, Monash University.
- \_\_\_\_\_ (1995). The politicisation of English: Part Two. The access test and the skilled migration program. *People and Place*, 3. Melbourne: Australian Forum for population studies, Monash University.
- Henning, G. (1987). *A guide to language testing*. Newbury House Publishers.
- Livingston, C., Castle, S. & Nations, J. (1989). Testing and curricular reform: One school's experience. *Educational Leadership*, 46 (7).
- Lynch, B. (1997). In search of the ethical test. *Language Testing*, 14 (3).
- Madaus, G. (1991). *Current trends in testing in USA*. Paper presented in the conference on Testing and Evaluation: Feedback Strategies for Improvement of Foreign Language Learning, February 4-5, Washington, DC: The National Foreign Language Center.
- McIntyre, A. (1984). *After virtue* (2<sup>nd</sup> ed.). Notre Dame, IN: University of Notre Dame Press.
- McNamara, T. F. (1991). Test dimensionality: IRT analysis of an ESP listening test. *Language Testing*, 8 (2), 45-65.
- \_\_\_\_\_ (1996). *Measuring second language performance*. Addison Wesley, Longman Limited.

- Messick, S.A. (1989). Validity. In Linn, R. L., *Educational measurement*. American Council on Education. MacMillan.
- Murphy, K. R. & Davidshofer, C. O. (1991). *Psychological testing: Principles and applications*. Prentice-Hall International, Inc.
- Pennycook, A. (1994). *The cultural politics of English as a world language*. London: Longman.
- Punch, M. (1994). Politics and ethics in qualitative research. In Denzin, N.K. & Lincoln, Y.S. (eds.), *Handbook of qualitative research*. Thousand Oaks, CA: Sage.
- Robinson, G. & Craver, J. (1989). *Assessing and grading student achievement*. Arlington, VA: Educational Research Service.
- Sanders, W. & Horn, S. (1995). Educational assessment reassessed. *Educational Policy Analysis Archive*, 3 (6).
- Seymour, R. T. (1988). Why plaintiffs council challenge tests, and how they can successfully challenge the theory of validity generalization. *Journal of Vocational Behavior*, 33, 331-364.
- Shavelson, J. R. & Webb, N. (1991). *Generalizability Theory: A primer*. Sage Publications.
- Shepard, L. (1991). Psychometricians' beliefs about learning. *Educational Researcher*, 20 (7), 2-9.
- Shohamy, E. (1993). The power of tests: The impact of language tests on teaching and learning. *National Foreign Language Center (NFLC) Occasional Papers*, June.
- \_\_\_\_\_ (1997). Testing methods, testing consequences: Are they ethical? Are they fair? *Language Testing*, 14 (3).
- Spolsky, B. (1995). *Measured Words: The development of objective language testing*. Oxford: OUP.
- \_\_\_\_\_ (1997). The ethics of gatekeeping tests: What have learned in a hundred years. *Language Testing*, 14 (3).
- Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *PHI DELTA KAPPAN*, 70 (9), EJ 388723.