Chapter 7

Situated Ethics in Language Assessment

Fred Davidson[1], *University of Illinois at Urbana-Champaign, and Antony John Kunnan*[1], *California State University at Los Angeles*

Introduction

We start this chapter on ethics and language assessment with a vignette.

Scenario 1. "The last test was not well designed."

Bellore University decided to develop its own Test of English, the medium of instruction at the university, for all its applicants for admission. The Dean of Graduate Studies assembled the English teaching faculty and asked them to develop an appropriate test in three weeks that was no more than two hours long and required only 10 minutes to grade. The faculty members were not happy about the timeline and the constraints and aired their concerns, but the Dean was very clear—the test had to be ready in three weeks and designed within the constraints mentioned.

The faculty members went about their task quickly. They designed a two-hour test that involved different language abilities, they experimented with new response formats, and the tasks required test takers to use English in a communicative way. Before they got around to designing scoring guidelines for these tasks, however, the three-week time limit was up. The faculty members went to the Dean and expressed their difficulty. The Dean was happy that the test was ready even though the scoring guidelines were not. The test was used the following week with the graders given 10 minutes to grade each paper although the graders had no uniform criteria for failure and success.

> When the results were ready, the Dean set an arbitrary passing cut score of 90 percent, and therefore only 10 percent of the test takers "passed" the test and were admitted to the University. But within a few weeks, faculty members started to complain that many students in the classes were not able to follow the lectures and instructions, read the textbooks and answer questions, and work on assignments and present project reports. The Dean then called the faculty and said, "The last test was not well designed; I'll give you another three weeks to design a new test. I want a better test on my desk but once again the test has to be only 2 hours long and take only 10 minutes to grade."

What can the faculty members do to improve this situation? How can they convince the Dean that more needs to be done to design an appropriate test? That more time and resources, among other things, need to be provided? These are questions at the heart of many steps in test development, design, administration, scoring, reporting, standard setting, and the social consequences of assessment, and assessment practice in general. One approach to improve assessment practice is to have professional standards coupled with responsible and ethical practice. Such a list of standards could be used to inform and convince decision makers, like the Dean, for a better system so that test development can be conducted in a more professional manner. And, such a list of standards could also be used as the basis for better assessment practice for all professionals, not only for those in large agencies who have their own in-house standards but also for classroom teachers who are responsible for much assessment. But key questions that are germane to this issue are: What are these professional standards? Where do they come from? On what basis can they be developed? Who develops them? And what does it mean to be professionally responsible and ethical, particularly when tests are used in multilingual and multicultural communities?

This chapter considers and presents our ideas for the development of an ethical milieu through a discussion of how ethical thinking developed in the field (through individual researchers and practitioners), the codified standards and codes through the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], National Council for Measurement in Education [NCME] 1999) and the Code of Ethics (International Language Testing Association [ILTA] 2000), and finally, through our reflections on a few vignettes intended to spark the dialogue necessary to begin discussions regarding the situated nature of applied ethics for the field.

Applied Ethics

Ethical theories offer ways to think about general abstract questions such as "What ought I to do?" "How ought I to live?" Applied ethics connects these theories to modern professions such as medicine, science, engineering, the environment, or education, and to specific pressing concerns such as abortion, the death penalty, or physician-assisted suicide. Both ethical theory and applied ethics can be religious or secular;[2] we restrict ourselves to secular ethics in this chapter.

The Kantian approach dictates that rational human beings ought to obey a "categorical imperative" that is universal and absolute or without conditions or exceptions. Examples of such rules exist in language assessment. For example, in the *Code of Ethics*, "language testers ought to have respect for the humanity and dignity of each test taker" (International Language Testing Association 2000, Principle 1), or "language testers shall not allow the misuse of their professional knowledge or skills, in so far as they are able" (ibid, Principle 4). Rules or principles could also be written at an applied level too, such as, "oral proficiency tests for employment should be conducted in a face-to-face manner," or "language assessment organizations ought to use at least two human raters to rate all essays."

A contrasting secular approach is outcomes-based: rules, principles, or evaluations of policies and/or actions are made on the basis of consequences. Also known as the consequentialist approach (Baron et al. 1997), this would focus on the consequential aspect of a test. For example, if results of oral proficiency tests for employment in a tape-mediated approach (as opposed to a face-to-face approach) are acceptable, then consequentialist ethicists would consider the tape-mediated test acceptable. Or, if one or two electronic raters can reproduce or correlate highly with the results of two human raters of essays, then this view would accept choosing the electronic raters over the human raters; or, it could be argued that a test does not need to be certified as a valid or fair test if no test takers, test score users, and stake-holders complain about the test.

Another secular approach to ethics is called virtue-based ethics: it exhorts all professionals to be virtuous and presumes that ethically responsible practice would naturally follow. The burden of ethical responsibility does not originate from an ethical system or standards or codes from some agency but rather from individual virtues of the particular professionals in that agency and in the profession. In this approach, the breakdown between religious and secular ethics is not distinct. Traditional virtues from religion-based ethics aremight include, for example, virtues of consciousness,

benevolence and self-restraint (from Buddhist ethics), humanity and goodness, rightness and duty, consideration and reciprocity, loyalty and commitment (from Chinese ethics), neighborly love, natural morality (from Judeo-Christian ethics), social and individual duties (from classic Hindu ethics), and obligatory acts such as charity, kindness and prayer (from Islamic ethics). The most recent concept of a care-based ethic (see Noddings 1984) or feminist ethics (Gilligan 1982) could be considered examples of virtue-based ethics. These virtues could be used to produce ethically responsible tests and assessment practice.

Yet another approach to ethics is relativism. Relativism takes the form "of a denial that any single moral code has universal validity" and that moral codes ought to be subject to "factors that are culturally and historically contingent" (Wong 1993, 442). Further, this position holds the view that it is wrong to pass judgment against anyone else who has different values or traditions or to try to force them to follow your own values or traditions.

Relativism has a certain appeal to language assessment. For example, a tester may present a defensible theoretical position in a lecture (in the morning) and then (in the afternoon) pursue an alternate but equally defensible practical course of action. In each situation, each position is defensible. In another example, a tester may feel strongly—and advocate to students and colleagues—that essay exams should be double-rated by human raters. He/She may later consult with an assessment institution that hopes to include an essay exam in its test battery but fears the cost of dual human rating. In this situation, the tester might agree to dual rating by a human and a computer if that is the only way to get an essay exam into the test battery.

Language testers are not alone in the comfort provided by relativism. Many ethicists find the relativist position a tenable one, particularly when a Kantian approach is the only alternative. For example, in response to the call for universal laws, the Dalai Lama (1999, 27) issued the following warning:

> No one should suppose it could ever be possible to devise a set of rules or laws to provide us with the answer to every ethical dilemma, even if we were to accept religion as the basis of morality. Such a formulaic approach could never hope to capture the richness and diversity of human experience. It would also give grounds for arguing that we are responsible only to the letter of those laws, rather than for our actions.

In sum, we believe that principles of ethical conduct will need to be developed through a "discourse ethic" developed by reasoned agreement (see Habermas 1982, for more on this view). Using this method, for

example, a "code of ethics" or "ethical standards" can be developed through rational debate among members of the language assessment profession in a community. Ethical decisions in language assessment are relative to given situations, but these decisions (one way or the other) also force us to consider categorical imperatives, to consider the outcomes of our actions, and to consider our own virtues. We have chosen to illustrate this dynamic through a series of vignettes, presented later.

Developing a Language Assessment Ethic

To develop a language assessment ethic, it is necessary to understand the last two decades of thinking in educational and language assessment that helped forge the ideas for the ethic.

Influence from Validity and Fairness

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, and postulated that test validity refers to the "appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores" (Messick 1989, 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, 20).

In this view of validity, Messick also explicitly advanced a critical role for value implications and social consequences and use, as part of test validity. This was the first time that values implications and social consequences were brought from the backroom (where test developers had conveniently ignored them) and included as part of test validity. This view has now been instantiated in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and the National Council for Measurement in Education 1999) and in many language textbooks and courses. Also, this view provided support for 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. 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.

After Messick's expanded validation view, the importance of fairness has also been better understood. The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and the National Council for Measurement in Education 1999, 73) acknowledged the critical importance of fairness as a goal in assessment by devoting a whole chapter to fairness for the first time. Its vision 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.

Further, the *Standards for Educational and Psychological Testing* specifically defined 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 a different perspective with the focus on three criteria for fair tests: 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 (2001) proposed a fairness framework that positions fairness as the ultimate goal in assessment and is inspired by two ethical principles: the principle of justice and the principle of beneficence.

Beginnings of a Language Assessment Ethic

According to Spolsky (1995), from the 1910s to the 1960s, social, economic, political, and personal concerns among key language assessment professionals in the United States and the United Kingdom dominated boardroom meetings and decisions. In the last two decades, though, 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 assessment, 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, Clapham and Wall (1995, 259) reviewed principles and standards but concluded "language testing still lacks any agreed standards by which language tests can be evaluated, compared, or selected." 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 that the ethicality of a test was to be judged. To offer guidance in this matter, a report of the *Task Force on Testing Standards* (International Language Testing Association 1995) was published, as well as more recently, a *Code of Ethics* (International Language Testing Association 2000) that lay out broad guidelines of how language assessment professionals should conduct themselves. This *Code of Ethics* could be used in all communities to develop, research, maintain, and evaluate tests as well as train professionals in test development and research methods with the usual caution, as the code may not be able to take into account local needs and concerns.

Standards and Codes

The ILTA Task Force on Testing Standards—1995

Description

Shortly after it was formed, The International Language Testing Association (ILTA) formed a *Task Force on Testing Standards* (TFTS). Its objective was to survey published world standards in language testing. The word 'standards' was originally conceived to mean standards of good practice—what ILTA later came to call a 'Code' of ethics, which was ultimately published and that we discuss below. The Task Force project discovered two other meanings of the word: standard as a widely accepted test, and standard as a particular level of expected performance. One hundred and ten documents or document sets were studied. Each became an entry in the Task Force report, describing full bibliographic information and providing a critical analysis of the document(s).[3]

Comments

The TFTS report is a valuable early effort in codification of practice in world language testing. It is logical that ILTA—as a new organization—felt that a survey was needed before it issued its own code.

The entries have become somewhat dated, as expected in the normal evolution of any academic enterprise. Regardless, we find this to be a unique document of bibliographic interest. The variety presented in the report is a good indication of the variation in world practice in language assessment.

AERA, APA, NCME Standards—1999

Description

Three U.S. organizations banded together in the 1950s to write a code of practice for measurement professionals. The *Standards for Educational and Psychological Testing* is now in its fourth edition under the sponsorship of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. This book-length document covers three areas of professional conduct: (1) test construction, evaluation and documentation; (2) fairness in testing; and (3) testing applications. Standards of practice are articulated for each area, and these are accompanied by commentary and discussion.[4]

Comments

The *Task Force on Testing Standards* report (International Language Testing Association 1995, 171) called this document "a powerful force, if not the key player, in the U.S. assessment guideline scene." Despite the U.S. authorship, the *Standards for Educational and Psychological Testing* are cited and utilized all over the world. They were influential in the discussions in ILTA as it sought to develop its own code of ethics, which we address next.

ILTA Code of Ethics (2000)

Description

In 2000, the ILTA adopted its own Code of Ethics. The opening commentary clarifies its intent: it is based on moral philosophy, and it seeks to outline principles of ethical conduct by language testers (International Language Testing Association 2000, 1). Nine ethical principles follow, and in the Code itself each Principle is accompanied by annotation. We present here the Principles themselves.

Language testers shall have respect for the humanity and dignity of each of their test takers. They shall provide them with the best possible professional consideration and shall respect all persons' needs, values, and cultures in the provision of their language testing service.

1. Language testers shall hold all information obtained in their professional capacity about test takers in confidence, and they shall use professional judgment in sharing such information.
2. Language testers should adhere to all relevant ethical principles embodied in national and international guidelines when undertaking any trial, experiment, treatment, or other research activity.
3. Language testers shall not allow the misuse of their professional knowledge or skills, in so far as they are able.
4. Language testers shall continue to develop their professional knowledge, sharing this knowledge with colleagues and other language professionals.
5. Language testers shall share the responsibility of upholding the integrity of the language testing profession.
6. Language testers in their societal roles shall strive to improve the quality of language testing, assessment, and teaching services; promote the just allocation of those services; and contribute to the education of society regarding language learning and language proficiency.
7. Language testers shall be mindful of their obligations to the society within which they work, while recognizing those obligations may on occasion conflict with their responsibilities to their test takers and to other stakeholders.
8. Language testers shall regularly consider the potential effects, both short and long term, on all stakeholders of their projects, reserving the right to withhold their professional services on the grounds of conscience.[5]

Comments

As language testers, we feel that the ILTA Code is a particularly relevant anchor. It represents a set of core values we engage on a frequent basis—daily at many times of the year. It is also the first effort by the international language testing community to speak to itself about such matters.[6]

Let us return to and interpret our opening vignette against the ILTA Code. In that story, the Bellore Dean directed the faculty to develop a quick-and-dirty language test. Once launched, the test drew criticism;

even though it 'passed' only ten percent of the candidates, those who passed still seemed unable to handle the workload in the target language courses. The complaints resulted in exactly the same directive: the Dean allowed a new test to be developed, but only under the same constraints.

Many of the ILTA principles can be engaged here, but to us, Principles 8 and 9 seem most relevant. The Bellore faculty hold obligations to their college and to their society, which obligations seem to conflict with responsibilities to the students (Principle 8). When the faculty complained, they were told to do the same thing all over again. Perhaps at this point, it is becoming a matter of conscience (Principle 9), and the faculty should consider refusing to work under the imposed constraints in test development and delivery. Ethical practice in language assessment often seems to involve one difficult choice against another. Fight or flee? Bend (maybe break) or resist?

We think there are middle grounds in these tensions. The Bellore faculty could make its concerns known, but agree to continue the quick-and-dirty assessment so long as Bellore allocates resources for future test improvement. The faculty could seek outside funding or contractual liaisons with language assessment institutions; Bellore could become a test-bed for new testing ideas marketed elsewhere in the world and negotiate a reduced-rate license to use those ideas in perpetuity. The university could look more closely at classroom assessment while at the same time allowing the quick-and-dirty test to continue: perhaps there is a weighted formula that gives more value to grade point average, class finals, or other data that may accompany incoming students.[7] Somehow, the faculty must seek to engage Principle 7, and ultimately improve the practices at Bellore. It may take time, but we hope it is feasible.

We close this chapter with a series of additional vignettes in language assessment ethics. Each story is intended to help us see the balances involved in language assessment practice.

Vignettes on Ethical Practice

Scenario 2. "Testing for No Apparent Reason"

Central University, State Campus (CUSC) is a large college in the mid-western United States. It is a land grant university, originally sited and funded by the federal government, but now dependent on money from its home state. It also relies on grant and contract funding from various research projects supervised by its highly touted faculty.

Over the years, certain departments have become more financially independent than others. Some units in the humanities (for example the Department of Foreign Languages) depend almost exclusively on state dollars. Other units receive state money but supplement it with indirect funds recovered from faculty grants and contracts. These other units have achieved a certain degree of autonomy in governance—they are almost, but not quite—colleges within CUSC.

One such unit is the Department of Corporate Practice (DCP), which offers several Masters Degree programs in Business Affairs. These MBA programs attract a large number of international students, for whom English is a foreign language. CUSC and DCP are in a constant, creative, diplomatic state of tension. On one side, CUSC sees itself as the guardian of all its graduate degrees: any degree conferred must adhere to campus guidelines and oversight. On the other side, the DCP feels it can best structure a competitive MBA that will place its graduates in key positions in the world of business.

This tension affects English as a Second Language (ESL). New CUSC international students take an ESL test on arrival on campus, and the results exempt them from or place them in campus-wide ESL service courses. The DCP has collaborated with the CUSC Intensive ESL program to create a series of business-oriented ESL courses. All international MBA students must take these business ESL courses—there is no exception; indeed, DCP prides itself on the English language training component for its international MBAs. However, CUSC requires those students to take the campus ESL test, even though it has no real effect on their placement. Regardless of the test score, these students wind up in the same business ESL courses.

Our Reflection

On the surface, this appears to be quite unfair and unethical. Principle 1 of the ILTA Code is at play: does it "respect . . . the humanity and dignity" of test takers to put them through a test that has no actual role in their studies? We do not question the DCP's right to require all its international MBAs to take a business English course; indeed, we applaud that. Rather, we question the need of the exam for this group. It cannot place them into the course, and it cannot exempt them. It seems to be a waste of time.

The matter may be a bit more complex however. DCP is one of few departments able to fund special-purpose ESL courses. Most academic departments simply do not have the money to do so or, perhaps, do not have sufficient numbers of international students to merit doing so. DCP

students receive specialized and intense instruction that propels them through their study. Most CUSC international students do not have that advantage: they take ESL classes that are heterogeneous by subject field.

Perhaps CUSC central administration has done the right thing. In requiring the international MBA students to take the test, the campus is anticipating and appeasing protest from students and faculty in units that cannot afford such a service—in fact, perhaps such protest has already happened. CUSC is requiring a sacrifice in lieu of a benefit. Principle 8 reminds us that language testing is just such a balancing act. The larger CUSC context may dictate a practice that, at a local level appears to make no sense.

Scenario 3. "What do you mean, 'It doesn't fit'? Let's make it fit!"

The nation of Exonumia has a long and rich linguistic tradition. Its citizens are routinely (at least) trilingual, speaking Exonumian as well as the languages of two adjacent nations: Verdigian and Kosoffian. In recent years, there has been a public demand for training in a language of world commerce, and after much public debate, Exonumian schools now offer Japanese, German, and English as subject foreign languages. By the end of secondary school, Exonumian students must pass a test in one of these languages, and in addition, students who matriculate at an Exonumian university are required to continue study in that language.

The Ministry of Education recently announced that all elementary and secondary education in Exonumia would be driven by a set of central 'performance descriptors,' carefully developed and articulated paragraphs of expected ability in various subjects at various school grade levels. The Ministry announces a quick implementation schedule: schools must demonstrate articulation with these new standards within two years.

Teachers of Japanese, German, and English held a hastily convened meeting of the CFLE (Council on Foreign Languages in Exonumia) in the capital, Obverse City. Among other heated topics, they asked themselves, "Do our various tests fit these new central performance expectations?" It was quickly apparent to the group that the answer was *no*. Most of the in-place language tests were developed, as one vocal participant stated, "totally ignorant" of the Ministry's new edicts.

After much debate, the CFLE decides to retrofit its exams to the central guidelines. They will simply survey all the test tasks, items, and even specifications and hunt for *any* remote match between *any* test and *some* word or phrase in the new central guidelines.

The CFLE decides to hold a news conference at the Grand Obverse Hotel, where it will announce its adherence to the new descriptors. This announcement will coincide with a private closed-door CFLE meeting in

which the foreign language educators will begin a much longer and more difficult task—to truly reform their testing practice, not only to match the new guidelines but also to improve its reliability and validity.

Our Reflection

Exonumian education is not alone. Teaching systems around the world are trying to implement or strengthen central control of teaching practice. And they are not alone in facing unreasonable timeframes for implementation. And, likewise, they are not alone in deciding to claim that their current assessment practice is already in tune with the central government's wishes. It seems that the testers are violating Principle 4 of the ILTA Code—they are stating that a particular test fits a particular group of language descriptors, even though the fit is not properly established.

Is that a polite fiction, or worse still, an outright lie? Probably not. It is probably true that the CFLC members can indeed find sufficient overlap with current practice to justify some evidence of content validation to the new guidelines. But this is a delicate point. The test-to-guideline map will be weak at best.

We find it interesting that the CFLE decided to both announce its match-up with the descriptors and at the same time convene a meeting to revise its assessment. Perhaps that is the best result of all. We find ourselves wondering: what is being done over in other teacher councils in Exonumia: math, science, history? Perhaps the CFLC is doing just exactly what the Ministry intended.

Scenario 4. "Just whom have I helped?"

One day, the editor of a language testing journal receives an unsigned letter. There is a post-it note on the letter, identifying the author as a "well-known member of our language testing community" and asking if the letter could be published anonymously:

"[name deleted] sent me an email about three years ago from [location deleted]. There was a lucrative corporate contract to develop language tests for several target languages in Africa. These tests would allow the company to find speakers of the various local languages who could truly function there at a high level—mastery of all the complex nuances of the various L2s. He was already in the thick of things, as is his character. It was a wonderful pan-glossic opportunity—we'd be working with NS informants. The test design problems were fascinating. He reminded me of a favor she had done for me and patronage he had paid me in years past. And the money was good, and quick, and the job seemed do-able.

He pushed every button. It worked; I packed a bag and went to join the advisory committee.

I am frankly proud of the technical quality of our team. We developed a clever and sophisticated test of proficiency in each of the languages in the region of Africa where this company would function. We did it by the book: full theoretical statement of construct, complete and rich specifications, a large task bank, piloting, revision of specs and tasks, a second pilot, additional changes, and finally four equivalent operational test forms. It was textbook test development. I felt alive and charged, every bit of my professional psyche tuned and working on the problem at hand. It would have been against my very nature to do otherwise (I say modestly).

I recently learned that the company did indeed roll into several African nations like a juggernaut. Its major product—[name deleted]—has now become the best-selling powdered children's milk in something like a sixth of the continent. It absolutely devastated its competitors. And it is falling prey to the very same problem as those competitors—users of the product water it down, children don't get enough nutrition, and the whole damn cycle is starting, all over again.

Today, in the cool light of relative distance, I feel guilty; causing malnutrition is also against my very nature. I wonder sometimes, if I had not participated, would the test have been written anyway? Would [name deleted] have convinced one of my colleagues to take my slot on that advisory committee? Just whom have I helped?"

Our Reflection

From time to time, language testers are asked to work on tests that have a complex, sinister, or even a deadly purpose. This story resonates with Principle 1, for ultimately the needs of the public are harmed. The writer of this narrative is of divided spirit. This corporate language test helped a company roll out a new product, and once the product was out, it was abused. Was the language tester partly responsible? Suppose the tester knew the product to be developed, knew that it was already being abused in some parts of the world, and knew the ethical dilemma. Suppose the tester refused a seat on the advisory committee. Would that really have made a difference? Worse still—suppose that the tester did not know the nature of the project until after the consultancy contract was signed, until after the advisory committees started, and until well into the workload. Could a member of the professional language assessment community withdraw, or must we all stick it out once our commitments are made?

Principle 4 says that language testers should not abuse professional knowledge or allow it to be abused, "in so far as they are able." It is sometimes quite hard to predict the effects of a major test and to know—during test development—just whom will be served.

Scenario 5. "A Tired Argument"

Professor Neville Mosely is an established member of the world language testing community. He is widely published, well-respected, and frequently consulted. His language testing course is legend, and it has figured keenly in the training of many language educators. There was an interesting discussion in his testing course not long ago. About halfway through the academic term, the group was working on item statistics and test development. The students were eager future foreign language teachers and researchers. The following classroom discussion transpired.

Prof Neville Mosely (NM): So, anyway, that's how norm-referencing works. You pilot a large number of items, selecting only those whose item statistics help you to form a normal or bell-shaped distribution. And because we know the area under the curve, we can state with confidence that a test-taker at a given score result is *at* or *above* a certain percentage of his or her peers, whom we call 'the norm group.' The meaning of the result (its reference) is that fact: the test score is widely interpreted as the percentage of people at or below that particular result and hence the test-taker's rank among the peer norm group. Many large-scale tests are built this way, like commercial measures, national tests, even tests in the corporate world. And language tests are no exception. There are many influential language tests that follow norm-referenced philosophy.

Cathy Miller (CM), a student: Professor Mosely, so are you saying that in norm-referencing, the most important thing is the shape of the total score distribution? Is that it?

NM: Yes, that's basically correct. So long as test development gives you that bell shaped curve, then you have achieved what you want to achieve.

CM: Doesn't it matter what you are measuring? What about the test content? I mean, aren't we supposed to have a table of specifications that sample across a desired domain, like theory or curriculum?

NM: Yes, you are supposed to have that also. But suppose that you sample some particular content area with just a couple of items, and suppose that those items do not produce good item statistics on the piloting—maybe the item-total correlation is extremely weak. Well, in most norm-referenced test development projects, those items will drop out through the normal evolution of the test, or possibly they will be altered. The

point is that the statistical criteria of quality form the gasoline to power the test development engine, not considerations of content.

CM: I'm sorry, but I am really bothered by that. Shouldn't testing be concerned with content?

NM: (under his breath): That's a tired argument.

CM: Sorry, what did you say, Professor Mosely?

NM: Um. I said: That's a good argument. Let's explore it a bit for the rest of the class hour.

Our Reflection

We would be the first to admit that strongly normative tests are frustrating; it is painful to see a test item wither and die because of weak item statistics.[8] Statistical determinism is a painful fact of the world.[9] The fact is that many social structures and advantages have a limited supply. The fact is that rank-based educational decision systems have evolved to accommodate that limitation. The fact is that normative assessment exists, but seeing it is like trying to see air. Likewise (perhaps), living without it is like trying to live without air. This is a very difficult lesson to learn, and it calls for a particular kind of vigilance.

We include this closing story not to raise the familiar ethical argument of norms versus criteria but to raise questions of ethics in the training of language testers, a point we see in Principle 5. Literature and scholarly attention to training is scant.[10] Our professor here stumbled. He recovered quickly, and we hope that Cathy was not offended and that the class proceeded to that wonderful discussion we know so well on the philosophical and ethical tensions between normative and criterion decision systems. Cathy's epiphany needs to be shared by the entire class, for only in such sharing do we equip language testers with a sense of ethics.

Phenomena at Play: Philosophy, Codification, and *Realpolitik*

We see three phenomena at play in ethics in language assessment. First, there are matters of philosophy. Ethics and morality have been, and continue to be, discussed by moral philosophers. This literature and conversation is of great value, and we encourage all language assessment professionals to seek its counsel. The second force at play is codification. Language assessment has formed into an international community, with

its own conferences, journals, and dialogues. Early on in the growth of this community, it looked to outside agents for ethical guidance (such as the APA/AERA/NCME). It has now developed the first of its own Codes; more are perhaps to come. Those who develop and use language tests can benefit from study of these internal community codes. The third and final force is political reality. No language test exists in a vacuum. It is not possible to develop an assessment and ignore the *realpolitik* of the test setting. We included the vignettes to illustrate that ethical practice is eventually a matter of political balance.

This chapter has discussed ethics and how ethical thinking has developed in the field. But moral philosophy, ethical codes and principles, and vignettes—whether fictional or real— cannot substitute for conversation about ethical problems.[11] Only through open discussion can we see ethics situated in the multilingual and multicultural communities in which language assessment operates. We hope that our chapter has sparked precisely this kind of dialogue. At the end of the day, there is but one constant: the myriad of ethical challenges faced by individual language testers. The key to success will always be the individual ethical practice of each language assessment professional, taken after careful personal consideration and after careful conversation with committed peers.

1 Both authors contributed equally to the concept and writing of this chapter.

2 See Singer (1991) for these secular as well as religion-based ethics.

3 The TFTS report is available at no charge via the ILTA Web site: http://www.dundee.ac.uk/languagestudies/ltest/ilta/publications.html

4 This document is not free, but it can be purchased from each of the member sponsor organizations; for example, via the APA Web site: http://www.apa.org/science/standards.html, which went into effect on June 1, 2003.

5 The Code is available free from the ILTA Web site; see note 3 above.

6 Similarly, Seebauer and Barry (2001) present ethics for scientists and engineers based on formal training required by the Accreditation Board of Engineering and Technology and the National Institutes of Health. Nagy (2000) renders the 102 Standards of the APA's Ethics Committee's Ethical Principles of Psychologists and Code of Conduct (which went into effect on June 1, 2003) into plain English.

7 Clearly, this solution requires an infrastructure of data reporting from teaching before the students reach Bellore.

8 For an example of this, see the Norton Pierce (1992) seminal article on the TOEFL reading exam.

9 For more on 'statistical determinism,' see Davidson (2000).

[10] There has been some, see Bailey and Brown (1995). In addition, there was a panel discussion on this point at the 2002 Ethics in Language Assessment Conference, Pasadena. See Kunnan (2002) for more discussions.

[11] This is also a major conclusion of Shohamy (2001). She advocates the view that the power of language tests can best be understood, survived, and perhaps reformed through open democratic discussion.