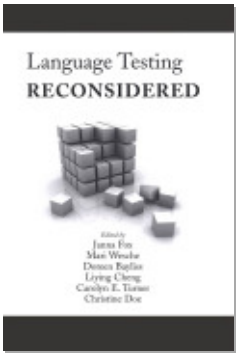
Language Testing Reconsidered

Fox, Janna, Wesche, Mari, Bayliss, Doreen, Cheng, Liying



Published by University of Ottawa Press

8 Tests as Power Tools: Looking Back, Looking Forward

Elana Shohamy
Tel Aviv University

Abstract

In this chapter I discuss current uses of language tests in education and society, arguing that tests have become primary tools used by policy makers to resolve and reform educational, political, and social problems. Specifically, I address two areas where this is happening: (1) in the realm of education, through the introduction of the No Child Left Behind tests in the USA, intended to reform education and resolve low school achievements; and (2) in the realm of society, through the increasing use of language tests for granting citizenship and thus, using tests to settle the complex set of issues related to migration. Relying on empirical research, I point to the length of time it takes immigrants to achieve academic language proficiency in schools and the continued role of L1; I argue that the use of such tests is unjust, unethical, and discriminatory and leads to marginalization and expulsion of people, suppression of diversity, and forced monolingualism. Further, these tests do not accurately represent current understanding of the language constructs of immigrants, who continue to negotiate and make meaning multilingually. I end the chapter with a call for the creation of language tests that are both in line with broader and more realistic language constructs, incorporate multilingualism, and multimodal realities, and also address the misuses of tests in order to lead to inclusion, participation, and recognition, especially given the ramifications of tests in creating de facto language policies.

Introduction

In this final chapter of *Language Testing Reconsidered*, I share my retrospection about language testing, based on my experiences in the field. In fact, this task has been instructive on a personal level as it forced me to take a hard look into my own motivations and complex relationship with language testing. Of course, I often wonder whether experience and seniority provide a sensible and creative perspective or whether they serve as devices that block and suppress thinking. Experience may provide an ability *to see*, but mostly using specific and pre-defined glasses. With this caveat in mind, I look back and offer insights from my experience, by providing my own narrative about language testing using my personal glasses, a narrative that has led me to develop a critical view of the state of the art in the field of language testing.

A Personal Narrative About Testing

Language testing for me, from the day I first started engaging with it, was about *reforming the world.* Being a victim of tests myself, I remembered very well my school days when tests were a hurdle, an *unpleasant* experience. It was tests that were responsible for turning the enjoyment and fun of learning into pain, tension, and a feeling of unfairness. Tests were often the source of anger, frustration, pressure, competition, and even humiliation. Coming out of tests was often accompanied by a feeling that my real knowledge could not be expressed. As a student in school, I recall often not understanding what tests were all about, why they were needed, and what their purpose was in the midst of rewarding and enjoyable learning experiences. Having to take a test often felt like betrayal. If learning is so meaningful, rewarding, and personal, why is it that it needs to be accompanied by the unpleasant events of *being tested*?

During my high school years, the testing experience was even more negative, as I was faced with *the big tests*, such as national tests at the end of high school or entrance tests to the university: it was clear that these tests could affect my whole future life. They could determine which university I would be accepted to, or whether I would be admitted to a university at all. High school was the period when tests replaced learning, as testing dominated both teaching and learning. I clearly recall how little meaningful learning took place during the last two years of high school, which were devoted almost exclusively to preparing for *the big tests.* The negative experiences of testing would spread to my home, where I was constantly being judged by my parents on the basis of my performance on the tests (often in relation to my other classmates), as if nothing else mattered. I do not remember being asked about "what I learned in school today" but rather "what grade I got on a given test." Tests became the sole criterion of success. Thinking about these issues now, I realize that students in schools rarely received any explanation of why tests were even needed, who benefited from them, and in what ways. But, as with all other school activities, students just complied. So although testing was an unpleasant experience, it was viewed as a necessary evil that was never questioned. Students were expected to conform to and accept, no questions asked, this integral part of schooling as just part of life.

Enrolling at the university for an advanced degree in language education and applied linguistics seemed to offer an excellent opportunity to delve deeply into the topic of testing, for no reason other than to better understand its mysteries, rationale, purposes, benefits, and costs. I was determined to understand the mysteries of testing and to possibly create better tests. I had a vision of tests as something different, perhaps without scores or without grades, without multiple choices that never made sense, and perhaps involving stimulating and exciting situations integrated in a productive way into learning.

Yet, when I became immersed in psychometrics, measurement, and statistics courses as requirements in graduate school, the focus was not on these types of issues. It was not about creating better tests from the experiential perspective or integrating testing with better learning, but rather it was all about formulas. It was understood that better tests would be achieved only if they became more reliable and valid by using sophisticated formulas and calculations, but there was nothing about *the testing experience*. In order to survive in graduate school, it was necessary to pass a large number of tests in all these measurement courses, similar to the types of tests administered in high school. Again, one had no choice but to comply: passing tests was clearly a demand in the field of testing, consisting at the time almost exclusively of men who knew math and statistics. These courses had little to do with real tests, with real people, or with real schools: they were not about experiences or consequences, learning or attitudes, but about creating tests that would be more accurate by following strict criteria. The tests discussed in graduate school belonged to the *big testing paradigm*, such as the TOEFL or the SAT, and not to tests given by teachers to students regularly in classes and schools as part of learning. The latter were simply viewed as irrelevant.

This was all in the era when language testing as an academic field had just begun to emerge. There were no *language* testing courses and only a very limited number of books on the topic. It was even before the journal *Language Testing* was born, about the time that the Canale and Swain model (1980) emerged and LTRC met for the first time.† One could find only occasional articles about language testing, mostly in journals such as *TESOL Quarterly* or the *Modern Language Journal*.

But it was also at this time that the FSI Oral Interview started spreading into academic communities, after being used exclusively by U.S. government agencies. It seemed like a very attractive test. It was the only test at the time in which people were asked to actually speak, to actually use language, in a face-to-face interaction. It was also attractive because the assessment of speaking was not based on scores and points of correct and incorrect words or phrases, but rather on a hierarchical scale that had broad definitions of language use.

My own research interest in those days, and for several years after, was directed at questions of test bias and method effects, addressing issues such as who was hurt and who benefited from certain language testing methods? How biased were certain testing methods? Should multiple choice or open-ended questions be used on language tests, and for whom? Should test questions be posed in L1 or L2? What other kinds of interactions could be used for testing oral proficiency to better reflect *real-life* oral interactions, in addition

†[Ed. note: See Bachman, Chapter 3, for a discussion of early models of language proficiency and early meetings of LTRC.]

to the oral interview, which was the single dominating method for assessing speaking at the time? Were specific tests or testing methods marginalizing certain students? Who were the most accurate raters for oral tests? What topics were being covered on oral tests, and were they appropriate for all students, for all populations?

It was my involvement in introducing a new oral English test battery, using multiple and varied types of oral interactions, into the Israeli secondary school system as part of a high school graduation test battery in English as a foreign language (Shohamy, Bejarano, and Reves, 1986), that made me wear different glasses. It was then that I began to see that none of the psychometric qualities of tests I had learned about in graduate school really mattered to policy makers. When it came to the Ministry of Education, the policy decision to introduce the new oral tests was based on different sets of calculations. The introduction of oral tests was not about accuracy of the results but rather about gaining a tool that could influence and control the teaching of languages in the classrooms of the whole nation. It was then that I began to understand the crucial role that tests played as instruments in political, educational, and social contexts. It was then that I observed how tests served as policy tools to be used by educational systems primarily to promote a variety of agendas and to exercise power and control. Specifically, in the case of the oral test, and in many other examples, the main purpose was to enforce the teaching of oral language in the classroom and the utilization of specific methods, materials, and tasks. Tests, I realized, were tools in the hands of policy makers used to impose and perpetuate specific agendas. I have since researched and discussed these issues in numerous publications, and more extensively in the book *The Power of Tests* (2001). Along the way, I also realized how detached we language testers are from the political, social, and educational realities in which our language tests operate and how blind we often are to the political realities and contexts.

Since then I have become very aware of the strong influence of tests on educational systems, of their ability to dictate teaching and learning methods, to define language knowledge, to determine what is considered correct language and what is not, to determine language hierarchies, to determine language priorities such as which languages count in given contexts, to perpetuate monolingual realities and suppress multilingualism, and to continue to perpetuate the imagined criterion of *the native speaker.*

These realizations led me to pursue a somewhat different direction in language testing, that is, a focus on the use of language tests in political, social, and educational contexts. I now believe that there is a need to study issues of test use and consequences and to focus on the impact and ramifications of tests and the motivations for introducing them. We need to study how language tests affect people, societies, teachers, teaching, students, schools, language policies, and language itself. We need to examine the ramifications of tests, their

uses, misuses, ethicality, power, biases, and the discrimination and language realities they create for certain groups and for nations, and we need to use a *critical language testing* perspective. All these topics fall under the theoretical legitimacy of Messick's (1994, 1996) work on the consequences and values of tests.

There are others in the field of language testing who ask similar questions: Lynch, Spolsky, Davies, Hamp-Lyons, McNamara, Norton, Elder, Cheng, Kunnan, Fulcher, to name just a few who examine different dimensions and perspectives on these issues. Because of this interest, language testing took a critical turn: it posed questions, it introduced doubts, it raised ethical issues, and it focused on fairness, responsibility, societies, and washback. It got us engaged in developing the ILTA (International Language Testing Association) *Code of Ethics* (2000) and discussing these issues at conferences on language testing and applied linguistics, in journal articles, and especially in the more recently founded journal, *Language Assessment Quarterly*.

Current Use of Tests in Education and Society

Yet, the current reality of tests, not only of *language* tests, is that they are given more power than ever, as they are widely used by governments, institutions, and central authorities world-wide. There is often a feeling that bureaucrats, educators, and political leaders have discovered an effective formula for providing the illusion that tests will solve all educational, political, and economic ills. Tests are currently used as the main instruments for educational reforms. Furthermore, in many countries language tests are used as gatekeepers to prevent the entry of unwanted people, such as immigrants and refugees, and thus to resolve national and international political issues. In the next part of this paper I will use a number of examples to show how tests are used for such purposes in education and society. I am at the same time claiming that it is also through the power of tests that progressive and open views can be negotiated and introduced.

Education

Education has always been a domain in which tests are used extensively for promoting and perpetuating multiple agendas. However, in the U.S., the No Child Left Behind (NCLB) Act (2002) provides a vivid illustration of how tests can be used aggressively and with only limited attempts to examine their consequences. The NCLB is anchored in law; it is sweeping in the sense that all students in U.S. public schools have to be tested, including recently arrived immigrants. Strong sanctions such as the closing of schools may be imposed on schools and teachers when students fail, and there are very limited possible repairs for schools whose students do not do well on these tests. Evans

and Hornberger (2005) show how the NCLB perpetuates negative effects on bilingual education in the U.S. Byrnes (2005) in the *Modern Language Journal*'s "Perspectives," includes articles that specifically show how the NCLB decreases students' motivations to learn foreign languages and schools' motivations to teach them, as they are not included in the NCLB; rather, students and schools invest in learning and teaching English, the language that is being tested. Like many other national tests that are imposed by governments and ministries of education, the NCLB is a powerful educational tool that creates *de facto* language policies. It perpetuates national languages as the only alternative, since all NCLB tests are offered in the dominant and hegemonic language, English, thus sending an indirect message regarding the irrelevance of other languages, especially languages of immigrants and indigenous groups.

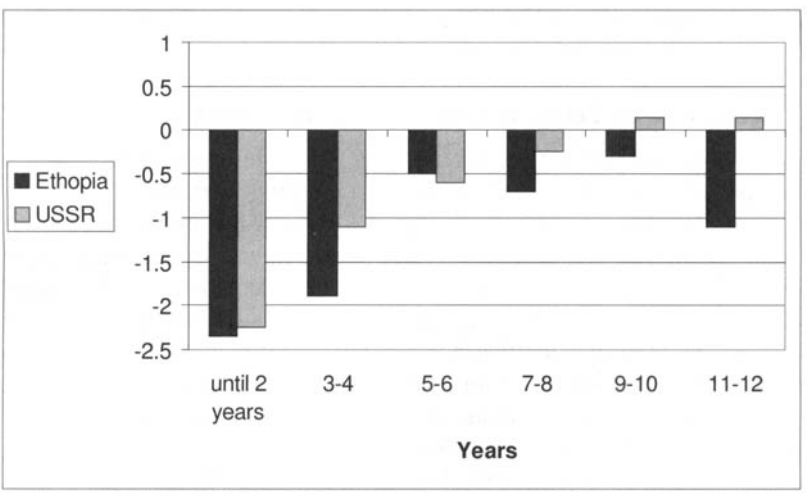


Figure 8.1: Ninth-grade Hebrew standard grades, according to years of residence

Further, while research has consistently shown that it takes a long time for immigrant students to reach equivalent levels of academic proficiency in the national language used for instruction in school, these tests expect immigrant students to reach levelss equivalent to those of students born into the language in an unrealistically short period of time, as they are tested through the medium of the new language a short time after immigration. Tests similar to those of the NCLB are administered in many nations nowadays. For example, Figures 8.1 and 8.2 are based on a national study conducted in Israel with Russian and Ethiopian immigrant students (Levin, Shohamy, and Spolsky, 2003). The graphs point to the length of time it takes immigrant students to acquire similar levels of academic achievement in Hebrew and mathematics in relation

to students of the same age who were born in Israel. Figure 8.1 demonstrates that it takes Russian immigrant students 9–10 years to achieve scores similar to those of native speakers. Ethiopian immigrants never achieve similar levels of academic proficiency.

Figure 8.2 shows similar results in mathematics; the graph illustrates, again, that it takes the Russian immigrants about 9–10 years to obtain scores similar to those of native speakers.

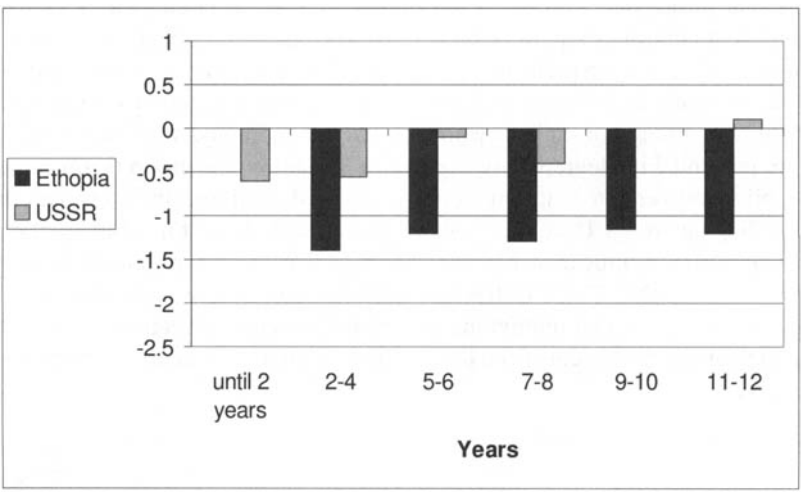


Figure 8.2: Eleventh-grade math standard grades, according to years of residence

The study measured these achievements in grades 5, 9, and 11. While slight variations occurred according to grade level, by and large the findings were the same showing consistently that immigrant students require 7, 9, or 11 years to attain equivalent levels of proficiency. It is especially revealing to focus on immigrant students from the former USSR, who generally arrive in Israel from educational contexts that provided them with high levels of mathematical knowledge and L1 literacy. Nevertheless, it takes them many years to reach levels of proficiency similar to those of the native speaker groups both in mathematics and language. The low achievement of the Ethiopian students is related to language mastery but may also involve other factors, as they do not come from an educational system that emphasizes advanced mathematics.

In spite of similar findings from other contexts (Thomas and Collier, 1997), indicating the length of time it takes immigrant students to acquire academic proficiency after they arrive, tests are often administered very soon after arrival, much before they are able to reach acceptable levels of proficiency, and poor results may lead to sanctions for students and for schools where the immigrants are enrolled. Clearly, administering these tests in an unfamiliar language

negatively affects the performance of immigrant students and may classify them, inaccurately and unfairly, as having limited academic potential.

The important role that language plays in the academic success of students can be demonstrated via a number of different types of analyses. One such method is the comparison of performance of students on academic tests that are presented in different forms with different accommodations. Figure 8.3 shows the performance of immigrant students on the same math test presented in two different forms, one a bilingual form in Hebrew and Russian, and one, to an equivalent control group, in Hebrew only. The results presented in Figure 8.3 show the difference in performance when the Russian immigrants were accommodated using the bilingual mathematics test version in comparison to a group of Russian immigrant students who received no accommodation, the questions were presented in Hebrew only. As Figure 8.3 shows, the group that received the bilingual version of the math test performed significantly better than the monolingual group. Thus, the group presented with the test in a bilingual format enjoyed a significant academic advantage over the control group (Hebrew presentation only). These differences persisted for up to 8 years after immigration, indicating that immigrants continued to create and construct meaning through utilizing elements from their L1 for a long time after arrival in the new country.

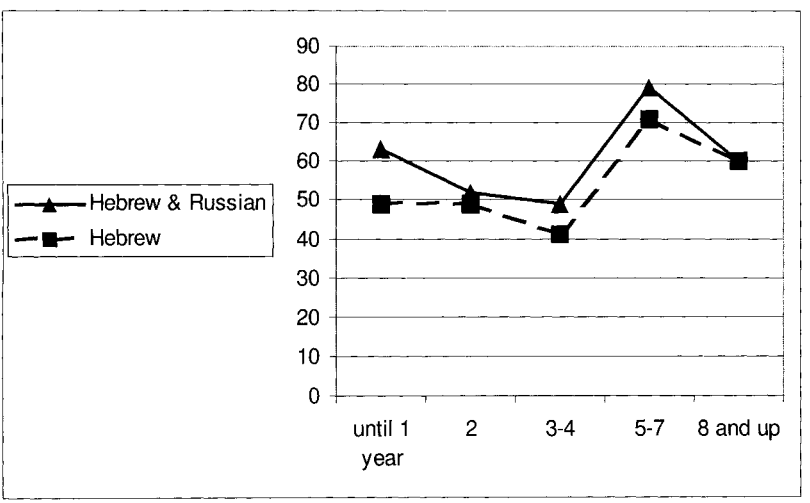


Figure 8.3: Math grades in monolingual vs. bilingual test versions

Given these results, one wonders what effect the presentation of tests with no such accommodations may have on the performance of immigrant students on tests such as those used in the NCLB program. In these tests immigrant

students are expected to demonstrate their academic knowledge exclusively through the new language. This ignores the fact that learning to function in a new national language is a long and complicated process, whereby meaning is created and constructed via multiple channels (Bialystok, 2001; Abedi, 2004; Solano-Flores and Trumbell, 2003).

The use of tests in such ways leads to a number of questions such as: Should tests be used as tools for educational reform? Does the administration of national tests lead to improved education and improved achievement? Are these tests a true reflection of academic language? Who benefits from such tests and who pays the price? What are the results of tests being used for? What are the *real* agendas behind introducing national tests? Is the use of such tests an attempt to create and perpetuate policies that are detached from what is known about learning? Finally, and of crucial importance here, what should the role of language testers be in supporting or objecting to the use of tests for such policies?

Society

The other example that illustrates current uses of tests to exercise power and control is in the area of national policies. The main example used here is the current use of language tests in a growing number of countries as a condition for right of entry, residence, and citizenship.

Many countries, especially in Europe, including Britain, Latvia, and the Netherlands, as well as the U.S., are administering language citizenship tests, and a number of other countries, such as Australia, are seriously exploring this possibility. The idea behind such tests is the belief that language proficiency, as exemplified through these language tests, is an expression of loyalty and patriotism and should be a requirement for residency, and especially for citizenship. It should be noted that citizenship is generally required for obtaining benefits and access to social security, health, education, and election rights. In the Netherlands, for example, language tests in Dutch are administered to prospective immigrants even before their arrival in the Netherlands as a requirement for immigration, in the name of social cohesion. The term *naturalization* in the U.S. context is revealing in terms of the ideology behind citizenship and the connections seen between language, patriotism, and negative views of diversity. In an era of globalization, relatively free borders, and trans-nationalism, language tests become the major tools for imposing societal uniformity and a way to gatekeep those whom nations view as undesirable.

In this case language tests serve as tools through which those who are not proficient in the hegemonic language are de-legitimized and marginalized in the places where they reside, in the workplace as well as in higher education, thus contributing to the creation of second-class citizens and often leading to

their expulsion. It is through language tests that such outcomes obtain legitimacy, as people are judged based on their language proficiency, and often at a level that is unrealistic for them to achieve.

Many questions arise with regard to the use of such citizenship tests: Why language? Why language tests? Why should language be considered a requirement of citizenship? How able are immigrants to acquire new languages at a certain stage of their lives or even to become successful in taking tests of a Western nature after being schooled elsewhere? Why is there a need to perpetuate national ideologies through language tests? (Are people possessions of states?) And why are language testers cooperating with policies that can lead to discrimination and gatekeeping?

Reactions and Actions

Given the above cases, and many others not discussed in this chapter, what is our role as language testers in being involved in supporting or rejecting such uses of tests given their potential consequences with regard to social justice? As is argued in Shohamy (2006), language tests are not neutral. Rather, they can determine language priorities and language hierarchies. They are capable of suppressing, eliminating, and marginalizing other languages; they can perpetuate national and hegemonic languages; they can define language knowledge and stipulate criteria for correctness (i.e., *the native variety*); and they can lead to the expulsion of people. Thus, as language testers we may at times find ourselves standing behind such acts and the use of language tests, uses, and consequences that may: lead to discrimination against those not proficient in the status languages, deny personal rights to those who cannot speak certain languages, marginalize other languages and other people who represent languages that are not tested, and perpetuate beliefs about language correctness, monolingualism, and the discrete boundaries of languages.

These are some of the questions and issues that need to be addressed by our profession: Are language tests always necessary? Do we, via our tests, contribute to an unequal world and harm social justice? Do our tests represent real language the way it is currently understood and used? Is the language we measure embedded in real and meaningful content? Do we provide legitimacy to discrimination, racism, xenophobia, and various political, social, economic, and personal agendas?

There is a need, I believe, for a deeper and more comprehensive understanding of language testing and its consequences, starting with those of us who work in this profession and then ensuring that educators, politicians, and others also gain such understanding. Given such an expanded understanding, we need to get engaged in a political debate about language awareness and language testing activism so we can influence those with better access to power

centers, educational systems, law making, and especially the implementation of language-related policies.

But for that to happen, languages testers need also to adopt a more open view of what we assess, i.e., *language*. We need to ask ourselves what construct of language we are working with and whether through the use of monolingual language tests that have very clear and discrete boundaries we are contributing to an unrealistic view of language that is more of an imagined construct than a reality. Do we stand behind tests that lead to narrow and unrealistic criteria of correctness, limited views of grammaticality, perpetuating specific accents, re-enforcing the construct of the native speaker variety and contributing to the suppression of multilingualism in societies — attitudes that may be promoted by nation states? In the case of immigrants, for example, it is known that they continue to construct meaning by using both L1 and L2 for a long time after entering the school system, a finding supported by Figure 8.3 in this chapter. This reality, which is not yet recognized by educational systems, is denied by the continuing use of monolingual tests. Users of languages, especially in multilingual societies, the dominant pattern nowadays, create meanings via hybrids and fusions. Thus, language tests need to reflect such multilingual realities, where meanings are created through mixes, hybrids, and fusions, where different codes are used for the purpose of communication and expression and where languages do not have such distinct boundaries as linguists have led us to believe. Furthermore, multiple codes exist within every language: that is, languages are also known to consist of elements such as visuals, pictures, images, music, art, graphs, and a variety of symbols with varied ways of "languaging" (Shohamy, 2006) and cross-linguistic language boundaries (Kress and van Leeuwen, 1996; Kress, 2003, 2001). As language testers we need to ask ourselves whether our tests reflect such "languages" and whether by overlooking such realities we are contributing to an unrealistic view of language, counter to its natural use in communication, and perpetuating artificial standards of correctness, homogeneity, purity, and other imagined normative features. As language testers we need to think of tests that will accommodate a broader construct of what we mean by language today. We need to consider multi-coded language tests that will represent multilingual and multimodal realities; this is especially relevant when English, a global language that is deeply embedded in most other languages in education, commerce, and academic life, is involved (Canagarajah, 2006). At the same time we need to be more aware of the misuses of tests, demonstrate their effects and ramifications, and protect the victims of *bad testing* in ways that can lead to more inclusion, participation, and recognition that is not just channeled through majority languages.

Given the above discussion, as language testers we need to better understand the powerful role that language tests play in creating *de facto* language policies, often operating covertly and implicitly, yet with strong ramifications

in terms of academic, personal, and human rights (Shohamy, 2006). Finally, regardless of the possible detrimental effects of tests in many domains, it is clear that tests are here to stay. That being the case, we need to begin exploring new ways through which these powerful and influential tools, given their consequences and ramifications, can be diverted towards creating positive, constructive, liberal, democratic, just, and negotiated tools that are capable of providing benefits, awards, privileges, language rights and freedom of speech. Tests need to be used not just to penalize *bad and impure languages* but to encourage the complex language varieties that are used among the diverse populations in this world and by extension to avoid the imposition of unrealistic criteria that serve only the privileged.

There is so much more to do; our work has only just begun ...