4 Language Assessment in Asia: Local, Regional or Global? A View from Israel

Bernard Spolsky
(Bar-Ilan University, Israel)

THE EXCESSIVE POWER OF EXAMINATIONS

The first examination system was started over 2000 years ago as an instrument of Imperial Chinese power, intended as a method of selecting the highest rank of government officials and replacing the patronage of powerful aristocrats by a process that aimed to measure academic skills. For two thousand years, it served as an elite winnowing procedure, fading only when the power of the Emperor was itself disappearing. As a method of selecting the very best, it did not have to worry about any possible wastage or unfairness or any of the other woes of mass testing. One of its main legacies has been the high status of testing in Asia, wherever Chinese cultural values were spread.

The tradition of testing was brought to Europe by the Jesuits (Ricci, 1942) and applied to the educational control of Jesuit schools (de La Salle, 1720). It provided a method of centralized supervision of classroom progress (Madaus, 1990), something that had previously been achieved in the local control of medieval schooling in Treviso in Northern Italy (Spolsky 2005; Swetz 1987), where the members of the local council went to school to test the pupils and decide then on the school master's salary; in the Jesuit schools in contrast, it was the school system that established the curriculum and the teachers who tested the pupils' progress, under the close supervision of the school principal.

Examinations had long been used at Oxford and Cambridge to test those students (not the majority) who wished to receive degrees that would qualify them to become teachers. The status of these elitist institutions and the qualities of those who succeeded in their

examinations encouraged the 19th century English politician Thomas Macaulay to propose applying what he called the "Chinese principle" to the selection of candidates for the potentially highly rewarding Indian Civil Service (successful administrators referred to as *nabobs* could expect to return to England with a fortune) (Thomas Babington Macaulay 1853, 1891). First adopted in 1858, the top 21 cadets were selected from 67 candidates who took examinations in Greek, Latin, English, French, German, Italian, Sanskrit, Arabic, mathematics, and natural and moral science. The examination system was later applied to selection for the English Civil Service, and by the end of the 19th century started to be used for quality control in primary and secondary schooling.

The switch from elite selection to control of mass education made clear both the power of high stakes testing and the narrowing effect it had on education: Latham (1877) was one of the first to complain about the "encroaching power" of examinations which were leading to a blurring of distinctions between liberal and technical education and encouraging the growth of crammers and cramming schools. Teaching, he said, was becoming subordinate to examination. A second major assault on examinations came a decade later, when a statistician (Edgeworth 1888, 1890) produced evidence of the inaccuracy (the "inevitable uncertainty") of the testing process, and thus its unfairness when used to make high stakes decisions in mass testing.

Both these warning were ignored, as in the 20th century centralized "objective" testing became increasingly powerful in Britain and US, and was subsequently spread throughout the world by the growth of a highly profitable psychometric industry. In the United States, otherwise intelligent politicians have been convinced that testing will solve the problems of under-financed schools. Everywhere, the testing business is booming: Pearson Education (part of a gigantic publishing industry) is starting to challenge more modest but highly profitable tests out of Cambridge and Princeton.

THE TESTING PARADOX

In this section, I want to describe briefly the nature of testing in general and language testing in particular. There is, as Taylor (2009)

remarks, a regrettably low level of "assessment literacy" among educational professionals, administrators, and the general lay public. There is a widespread tendency to assume that the tester, whether visualized as a distinguished professor in his academic garb or as a white coated technician feeding data into her computer, knows best. The naïveté of the journalists who write about testing is legion: I recall headlines in Isrtaeli papers complaining that half of the candidates scored below the average! My keenest memory was sitting with a Minister of Education complaining about the lack of psychometric sophistication in the agency responsible for the school leaving examination: I don't see the problem, he said: testing is easy: I used to be a teacher: you just make up some questions and the boys who get them right know the material.

It was Edgeworth who put the cat among the pigeons when he demonstrated by statistical calculation the errors built into traditional testing: the mistakes made by graders, the effect of candidate's health, the uneven results of question selection, the standards of markers varying according to personal whim, fatigue, or health. This was confirmed a half-century later by studies of Sir Philip Hartog (Hartog, Ballard, Gurrey, Hamley, & Smith 1941; Hartog & Rhodes 1935, 1936) who showed the variations in the marks given by different examiners and by the same examiner on different occasions. The field of psychometrics, developing at the end of the 19th and first half of the 20th centuries, took as its main task to study the measurement of human abilities and attempted to obtain something like scientific precision. Its earliest work was to develop more or less objective measures, typified by the multiple choice examination with enough items to produce easily calculable statistical estimates of reliability and validity; it proceeded further to devise methods to improve the reliability and validity of subjective marking; and most recently has turned to detailed consideration of the interpretation and use of test results.

For the layperson, a test question is a question. But, as Searle (Searle, 1969) noted, the normal condition for a well-formed question is that the questioner want to obtain information which he or she doesn't have but can reasonable expect the interlocutor to know and be willing to answer. Of course, in the case of the examination question, the questioner is the one assumed to know the answer. So it makes more sense to consider examination questions as requests to perform; the marking or grading

then becomes an evaluation of the performance.

In some cases, the performance is the actual activity it is intended to predict. The most obvious example is a driving test, where the candidate drives a car and the driving examiner determines that there has been no mistake. A performance test can also be verbalized: an interesting early example is the British Navy's examination for lieutenants, who were interrogated by senior captains as to how they would perform in different circumstances. A test based on a defined curriculum (for instance, knowledge of a pre-determined list of words in a foreign language) is also a fairly obvious performance test, but it already presents some problems. One is the sampling issue. Assume we have taught a hundred words: how many do we need to test to satisfy ourselves that the whole list is known? But assume we test only a sample: how do we make sure it includes words of equivalent difficulty? The second question concerns the method of testing. Read (2000) has made clear how many different ways there are of "knowing" a word: recognizing a definition, recognizing a translation, being able to give a translation or definition, using it in a sentence with a blank, making up a sentence with it… The list goes on: which is relevant to our purpose?

Which brings us to the central issue now being faced in testing theory: what is the purpose (interpretation, meaning) of the test and its results? I have puzzled over this question for a long time: Spolsky (1973, 1985) each dealt with the question "What does it mean to know a language" the first going on to ask "how do you get someone to perform his competence?" and the second asking about the theoretical basis of language testing. Come back to the naval lieutenant's examination: how many of the myriad names of parts and ropes of ship should he know? How many naval emergencies should be able to describe and save? Or in language, do we want to test vocabulary knowledge, or grammar, or pronunciation, or pragmatics. In fact, which of the fifty or so different scales ("addressing audiences, asking for clarification, coherence, communication strategies, communicative activities" are the first five) presented in the Common European Framework (Council of Europe, 2001) do we need to include? It all depends, obviously, on the purpose of the test. But one thing is obvious: given the complexity of the human abilities we are trying to measure, it is clear that a test for one purpose (selecting the student best qualified to represent our school in a debate

contest, for instance) will not serve for another (deciding which of our pupils need further drill in spelling).

I have argued that good tests must be

- Fair (reliable, unbiased)
- Relevant (related to purpose and use)
- Skeptical (cautious in interpretation, allowing for error)
- Efficient (The importance of the decision for the individual controls the cost)

Let me describe each of these in turn. For more traditional testing, it was often enough that a test be "felt fair", [1] that is to say that the public and the test takers felt that the test was being fairly set and administered. Traditional tests were prepared by professors or teachers in the field being tested and checked before use by experienced examiners. The problems emerged when such tests were exposed to more careful psychometric examination. One principle that is applied in standardized testing is to pre-test items: the questions that will be used are given in advance to an equivalent group of candidates, and the responses analyzed to see how difficult the items are in practice and how well they discriminate among the population.[2] In a multiple choice test, one is interested to see what proportion of candidates chose each distractor. At this stage (or after a test administration), one might further check for test bias by looking to see if various sectors of the population (males, females, minority students, new immigrants, for instance) did especially well or badly.

The next interesting question is how hard or easy was this test form compared with one used previously. In traditional examinations, this process was carried out by inspection by experienced examiners; in psychometric testing, anchor items are inserted in large tests and any

[1] This was a phrase I first heard from the Secretary of the University of Cambridge Local Examinations Syndicate.

[2] Difficulty is usually assessed as the percentage who answers correctly; the real life measures often turn out to be quite different from even experienced teachers' guesses. Discrimination is assessed by asking what proportion of students who scored best on the test as a whole answered this question correctly.

variation on these shows whether the cohorts are matched.

It is also considered important to control for possible cheating. Examination papers (and sometimes examiners and examination printers) are often locked up; the candidates are proctored by reliable people other than their teachers; candidates are watched for their behavior (cell-phones have produced a new problem). There are also statistical techniques to check for cheating (for instance, a student answers correctly an item that should be too difficult).

Making sure a test is reliable seems easy, but in fact it is not. When Cambridge was trying to have its English test compared to TOEFL (Davidson & Bachman, 1990), the comparison broke down because the researchers could not establish the technical reliability of the British measure. In China, in spite of rigorous procedures (the CET is given at the same moment all across the country, and security agencies are responsible for distributing the test), cheating is known to be prevalent. In Israel, one of the main reasons that the universities started their own test in 1981 for admission was the unwillingness of the Ministry of Education to tell them which schools were known to be actively involved in cheating.

The second set of qualities is relevance: how do we know the test is measuring what we want to measure? There are various ways of assessing the validity. One that was (and still is) commonly used is that this new test produces much the same results as an older one. When TOEFL was first used, its validity was justified by comparison with the older Michigan Lado test; it was considered preferable because new forms were produced every time, and so one did not have to worry about candidates memorizing the questions. A second is that its various parts more or less correlate (vocabulary, grammar, etc) but not so closely as to make them all a test of the same thing. A third is that the items are logically related to the test purpose: it is reasonable to assume that language proficiency includes vocabulary and grammar and comprehension and writing. A fourth is that the results appear to predict the desired behavior – the Scholastic Achievement Test was considered a fair predictor of first year results at university. All of these suggest the relevance and validity of a test, but there is no precise measure (the NITE university entrance test seems to account for about 40% of the variance in first year results, for so many other things can happen to students between the examination and the end of the first year of study).

Validity assessment is therefore complex and difficult, and is seriously compromised when a test is used for purposes other than it was intended, e.g., using a language test to predict academic success, or to hire employees.

Taking into account these two factors, which help explain the "inevitable uncertainty" of examinations, you will see why I put skepticism about interpreting the results next. Edgeworth showed (and statistical analysis continues to reveal) the imprecision of results. Psychometrists always assume there is a *standard error of measurement*, a calculation of the degrees of confidence for an individual score. This produces serious problems in setting a pass mark, particularly towards the middle of a ranked population. In the best of cases, with carefully prepared and pretested examinations, the error is likely to be several points, so care is needed in making a difference between a score of 50 and 55, so instance.[3] You can be reasonably confident that the top 10% of candidates have passed, and that the worst 10% have failed, but other decisions in large-scale tests are much more doubtful. There are then serious ethical questions in using precise examination scores as gatekeeping instruments.

Finally, there is the question of efficiency. Here of course the paradox is obvious – the cheaper (shorter, more easily administered) a test is, the less reliable it is going to be. A colleague of mine used to argue that the best test for language proficiency was a simple vocabulary test, until candidates knew you planned to test that way; at that point, it became a measure of memory and not language. Again, there is a paradox. Preparing for a test by practicing or memorizing expected test items can raise scores,[4] but without necessarily adding to the relevant proficiency. But as the stakes of test results become higher (we cannot get into a university without a good score, or we cannot get a job without it), so the impulse for preparation (cramming) increases, meaning that teaching becomes subordinated to testing. This of course was Latham's

[3] Israeli courts have just ruled that Bagrut papers that fall into this area must be remarked or granted passes.

[4] For many years, Educational Testing Service denied that test preparation firms like Kaplan produced any benefit for candidates, but finally they admitted the effect by offering their own test preparation services.

complaint a century ago, and it is just as true in Israel today as high schools devote much of their last few months before examinations to test preparation.

This has turned out also to be one of the major issues raised against the current US obsession with national standardized testing embodied in the No Child Left Behind Act. A quick word of background might help. The US Constitution is clear that responsibility for education devolves on state governments and not on the Federal Government. From time to time, the Federal Government has wanted to influence education, and the method adopted has been to make funds available for teaching certain curricular areas (science and languages after Sputnik, for example) or to encourage training in certain areas (teaching or engineering or graduate education). The concern has always been how to assure accountability, how to make sure that funds have been used usefully. The Act was passed in 2001, and followed discussions in the years before with state governors about how to improve education. It was based on the notion of granting funds to schools provided that they used state wide tests and showed evidence of improvement in results every year. These tests were of course limited to core fields like mathematics and reading, and one result was the reduction of teaching of foreign languages, art, music, etc. Another effect was the complexity of dealing with immigrant pupils, who have been required to take the same tests. There is a bitter controversy between supporters of the Act who claim that there have been overall improvements in reading scores, and opponents who argue that the improvements reported include years before the Act had any effect.

But the general tendency is clear – many politicians and educational administrators are convinced that centralized testing is an appropriate way to improve educational efficiency.

LOCAL, REGIONAL, NATIONAL, OR GLOBAL TESTING

This brings me to a central issue of this series of featured presentations, the issue of localization or centralization. It is of course a political issue. There are good pragmatic reasons why large centralized testing firms can produce better tests than local teachers or schools.

When Educational Testing Service took over the Test of English as a Foreign Language (Spolsky, 1995), it had the professional staff and knowledge and computers to produce a test more quickly and efficiently. In the first year of TOEFL, when it was still an independent concern, the director and associate director spent most of their time rewriting the items submitted by the carefully selected university English teachers, and the actual administration (test printing, distribution, registration, notification of results) was to be left to ETS. ETS had no problem fitting the process into its well-developed and efficient preparation and administration machinery, and once some minor problems were solved,[5] the growing demand for international English testing meant that TOEFL was for several decades the best moneymaker for ETS. Over the years, there were changes in the examination but only under external pressure,[6] and it was only four decades later that major revision was undertaken. By this time, of course, the competition had also grown strong – the Cambridge ESOL tests were well established, and the largest educational publisher in the world was about to launch the Pearson English test.[7]

The fact that English testing is big business means that globalized testing can be profitable, and the tests are likely to be more reliable and more efficient than local tests. Similarly, we will see in the next section, where I will sketch some aspects of Israeli English language testing, there is good reason to professionalize and even industrialize the production of standard tests on a national level rather than leaving it to teachers to make up their own class tests.

---

[5] To keep down expenses, a number of test modules such as speaking and writing had been omitted from the original design, but the unpredicted expense that led to losses in the first few years was mailing test results to any US university that the candidate requested.

[6] The director of TOEFL was for many years an administrator and not a tester. The ETS policy of not reinvesting TOEFL profits in the test contrasts with that of UCLES, which once it had recovered from the shock of the report of the comparison study (Davidson & Bachman 1990) undertook a program of building up testing staff and constantly revising its test (Spolsky 2004).

[7] As I write this paper, I see that Oxford is about to announce its own competing battery of computerized English tests.

Here though is the central paradox: just as the quality of education depends finally on the ability of a teacher to develop methods (and content) appropriate to her own pupils, so every attempt to centralize testing narrows the possibility for adaptation to local needs and goals. Again, it is a political or philosophical question rather than a purely educational one; the assumption that the central authority always knows best is highly questionable.

SOME EXAMPLES FROM ISRAEL

I have taken longer than perhaps I ought in setting out general questions and only now apply this to some examples of English language testing in Israel. First a few background points will help. The Israeli educational system was set up as a state system with independence in 1948: previously, the British Mandatory government had left education to the Arab and Jewish communities. In the last year before the end of the Mandate, a committee had met regularly to discuss language policy for the new system. There were several proposals, including one to use English as language of instruction in secondary schools. In the last meeting, assuming no doubt that there would be two states, the committee came up with a formula based on the language rights approaches of the Treaty of Versailles in 1920: each school would use either Arabic or Hebrew as medium of instruction depending on the majority of its pupils, and would teach the other language and English as additional languages. In practice, this policy applied only in Israel: the Arab areas were taken over by Egypt and Jordan, which applied their own Arabic only policy.[8] The Israeli State[9] schools then used either Hebrew or Arabic as medium of instruction for the full 12 years of schooling, teaching Arabic or Hebrew, French and English as additional languages. As the years went by, English became increasingly valued as a language for advanced education (the

[8] Jordan allowed Christian missionary schools to teach in their metropolitan language until 1962 when all were required to teach in Arabic.

[9] Ultra-orthodox schools were funded by the government but established their own curricula. Many used Yiddish rather than Hebrew as medium, and few taught English.

universities did all their teaching in Hebrew, but students had to read a lot in English), for business and for tourism.

The English examination given as part of the final high school examination, called the *Bagrut* [10] became a significant feature and a major challenge for those wishing entry to university. [11] The examination, like the curriculum for English, was traditional – some grammar questions, some reading comprehension, some questions on the set texts (including a Shakespeare play), and an essay. It was set by a group of teachers under the direction of the Inspector and monitored by a university teacher who would declare after inspection that it was harder or easier than last year. The Ministry contracted with a firm to have the test printed, distributed to schools, and hand marked by teachers hired for the purpose. No psychometric analyses were conducted or available. One of the conveniences of the system was the extra power it gave to the Ministry, first to use the examination for control of the school curriculum, second to issue its own optimistic interpretations of the overall results, and third to deal with otherwise embarrassing situations.

Let me give an example. The practice was early introduced that the final grade for a subject consisted of whatever grade the scoring process produced averaged with a grade reported by the subject teacher, who was supposed to take into account quality of class work and judgement of elements not included in the formal examination. At one stage, there were complaints that some teachers were over-generous, and therefore a technique was introduced that the teacher mark could not be too much higher than the exam mark.[12] But there was another use of the system. When we started a study of academic achievements of immigrant pupils in Israeli schools, we asked the Ministry for any comparative data it had, and were shown the *Bagrut* scores that had been presented to the Knesset Committee on Education. These showed that Russian

[10] Bagrut means "maturity", the term used for the examination in many Eastern European countries.

[11] When I first arrived in Israel in the late 1950s, it was normal for families to hire a private tutor to help prepare 12th grade students for the English *Bagrut.*

[12] There were also complaints that some elite schools were too strict in their teacher marks.

immigrants were scoring higher than native Hebrew speakers on tests of Hebrew and History. We then learned that these students did not take the examination, but that teacher marks were being reported as though they included the examination mark, thus seeming to show the immigrants were adapting well and rapidly. Our own subsequent research using reliable tests showed that in fact it took immigrants about six years in the country to catch up with native speakers (Levin, Shohamy, & Spolsky, 2003).

Over the years, minor changes in the English examination were made. In the 1960s, a new "communicative" curriculum dropped literature testing but included cloze items and an aural comprehension section (the passage was radio broadcast at the time of the examination). In the 1980s, a major attempt was used to modify curricular emphasis by developing an effective method of oral testing (Shohamy, Reves, & Bejerano, 1986). This appeared to have some influence, but the Ministry was not prepared to meet the costs of the number of examiners needed to assure reliability. Most recently, bribed by the offer of funding, the English Inspectorate is struggling to apply a discredited fifty-year old psychological theory to the testing (and so teaching) of literature.

In the meantime, a new problem had arisen, as the universities became increasingly dissatisfied with the examination and its usefulness for admission purposes. They had at the end of some years of complaining two major issues: a request that the Ministry inform them of those schools where it knew there was regular teacher-assisted cheating during test administration, and a request that the Ministry report examination results within two years of administration.[13] The Ministry was unable to unwilling to make these guarantees, so that in 1981, the heads of the Israeli universities set up the National Institute for Testing and Evaluation, staffed by professionals trained in psychometrics, to provide a reliable and valid instrument for admission to universities. The Psychometric Entrance Test consists of a battery of instruments (including Hebrew or Arabic proficiency, English language proficiency, mathematics and other cognitive skills) that is administered

[13] I was told of one case where the results were not available until the candidate had completed four years of army service and another three years at university: at that point, he was informed he had failed his *Bagrut*!

five times a year and is available also in Arabic, French, Russian and Spanish). Needless to say, the test endeavors to meet all the criteria I listed earlier in the paper, and aims only to predict successful completion of university studies. It is thus relieved of the confusion of the *Bagrut* which mainly serves the purpose of effecting Ministry control of the curriculum. Special attention is paid to the problems of minority students (Elliot Turvall, Bronner, Kennet-Cohen, & Oren, 2008) and of students with disabilities (Cohen, Ben-Simon, Moshinsky, & Eitan, 2008; Oren & Even, 2005), and the Institute produces several research reports each year.

The Ministry is far from happy with the existence of the Psychometric Entrance Test, which challenges its power. It criticizes it as elitist, in spite of the evidence that it serves the needs of candidates much better with its regular administration and test accommodations, particularly considering that many of those taking the examination do so after some years of compulsory army service. It continues to defend its inadequate testing service.[14] It also tries to pressure the universities to rely on *Bagrut* scores in admission, at least in the less competitive programs.

There are however some signs of progress. For some years, the Ministry has been conducting tests of progress in various subjects and levels. The task was contracted to the National Institute of Testing and Evaluation, which produced (as would be expected) highly professional tests within the limitations set. The National Authority for Measurement and Evaluation in Education (RAMA) has recently assumed responsibility for the *Meitzav* examinations. RAMA was established as an independent governmental body, reporting to the Minister of Education, "to lead and guide the Israeli education system regarding all aspects of measurement and evaluation." Headed by a scholar with serious psychometric qualifications (she held a senior research position at Educational Testing Service), it will no doubt maintain the quality of these standardized tests. But so far, RAMA has not been given any responsibility for *Bagrut*, which remains the preserve of Ministry officials with little or any sophistication in the field of assessment.

[14] The Ministry responded to recent complaints that *Bagrut* examinations papers were not being proofread by asserting that they had years of experience in giving examinations.

THE PARADOX OF EDUCATIONAL ASSESSMENT

Israel has not been spared from the "encroaching power" of examinations, nor has the presence of a core of assessment scholars[15] avoided the political attractiveness of using centralized examinations for ill-conceived efforts to remedy weaknesses in an educational system. Tests, like guns and medicines, have useful functions, but like them, they can easily be misused and their dangers can be ignored. Used modestly and skeptically by a teacher, they are an excellent way to assist in teaching. Used immoderately and ambitiously by central educational administrations, they are a sure way to blunt teaching initiative and narrow the kind of education offered. Sold irresponsibly by national or international businesses, they are a sure way to make money. There remain fundamental questions as to the extent with which human abilities can be measured or reduced to unidimensional scale, so that they serve as Procrustean beds that deform and distort whatever they are applied to.

REFERENCES

Cohen, Y., Ben-Simon, A., Moshinsky, A., & Eitan, M. (2008). *Computer Based Testing (CBT) in the service of test accommodations*. Jerusalem: National Institute of Testing and Evaluation.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching assessment*. Cambridge: Cambridge University Press.

Davidson, F., & Bachman, L. (1990). The Cambridge-TOEFL comparability study: An example of the cross-national comparison of language tests. *AILA Review, 7*, 24-45.

de La Salle, Saint Jean-Baptists. (1720). *Conduite des Ecoles chrétiennes*. Avignon: C. Chastanier.

[15] For example, the Academic Committee on Research in Language Testing has organized annual professional and scholarly meetings for over twenty years.

Edgeworth, F. Y. (1888). The statistics of examinations. *Journal of the Royal Statistical Society, 51*, 599-635.

Edgeworth, F. Y. (1890). The element of chance in competitive examinations. *Journal of the Royal Statistical Society, 53*, 644-663.

Turvall, E., Bronner, S., Kennet-Cohen, T., & Oren, C. (2008). *Fairness in the higher education admissions procedure: The psychometric entrance test in Arabic*. Jerusalem: National Institute of Testing and Evaluation.

Hartog, P., Ballard, P. B., Gurrey, P., Hamley, H. R., & Smith, C. E. (1941). *The marking of English essays*. London: MacMillan.

Hartog, P., & Rhodes, E. C. (1935). *An examination of examinations, being a summary of investigations on comparison of marks allotted to examination scripts by independent examiners and boards of examiners, together with a section on viva voce examinations*. London: Macmillan.

Hartog, P., & Rhodes, E. C. (1936). *The marks of examiners, being a comparison of marks allotted to examination scripts by independent examiners and boards of examiners, together with a section on viva voce examinations*. London: Macmillan .

Latham, H. (1877). *On the action of examinations considered as a means of selection*. Cambridge: Deighton, Bell and Company.

Levin, T., Shohamy, E., & Spolsky, B. (2003). *Academic achievements of immigrants in schools: Report to the Ministry of Education*. Tel Aviv: University of Tel Aviv.

Macaulay, T. B. (1853). *Speeches, parliamentary and miscellaneous*. London: Henry Vizetelly.

Macaulay, T. B. (1891). *The works of lord Macaulay*. New York: Longmans.

Madaus, G. P. (1990). *Testing as a social technology*. Boston: Boston College.

Oren, C, & Even, A. (2005). *The fairness and validity of the higher education selection system for students with disabilities*. Jerusalem: National Institute for Testing and Evaluation.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Ricci, M. (1942). *China in the sixteenth century: the journals of Matthew Ricci, 1583-1610* (Louis. J. Gallagher, Trans.). New York: Random House.

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge: Cambridge University Press.

Shohamy, E., Reves, T., & Bejerano, Y. (1986). Introducing a new comprehensive test of oral proficiency. *English Language Teaching Journal, 40*(3), 212-220.

Spolsky, B. (1973). What does it mean to know a language, or how do you get someone to perform his competence? In J. W. Jr. Oller & J. C. Richards (Eds.), *Focus on the learner: Pragmatic perspectives for the language teacher* (pp. 164-176). Rowley, Mass.: Newbury House.

Spolsky, B. (1985). What does it mean to know how to use a language: An essay on the theoretical basis of language testing. *Language Testing, 2*, 180-191.

Spolsky, B. (1995). *Measured words: The development of objective language testing*. Oxford: Oxford University Press.

Spolsky, B. (2004). Review of continuity and innovation: Revising the Cambridge proficiency in English examination. *English Language Teaching Journal, 58*(3), 1913-2002.

Spolsky, B. (2005). The Treviso language test: Some principles. *Quaderni di Ricerca del Cli*, 1-8.

Swetz, F. (1987). *Capitalism and arithmetic: The new math of the 15th century, including the full text of the Treviso arithmetic of 1478* (David Eugene Smith, Trans.). La Salle Ill: Open Court.

Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics, 29*, 21-36.