Russian Language Journal, Vol. 55, 2001-2005

The state of the art in language assessment: Notes for the third millennium

Bernard Spolsky

In the 2000 years during which human abilities have been assessed formally, tests and examinations have grown more powerful. A century ago, critics launched a strong attack on examinations, citing their "inevitable uncertainty," but a growing testing industry and governmental cries for "accountability" have managed a stubborn defense. More recently, appreciation of the complexity of notions such as "language proficiency" and acceptance of the resulting impossibility of finding a single measure of those notions have led testing experts to a realization that assessing language knowledge is multipart and intricate—and more likely to be served by profiles than by simple scores.

Seeking simplicity: The first 2000 years

If we accept as the beginning of formal testing the development of examinations on classical Confucian doctrine during the Han Dynasty (201 BCE to 8 CE), we have 2000 years of history from which to derive our understanding of testing—and on which to base our assessment of the current state of the art.

The Chinese examinations were designed as a method of selecting senior civil servants in place of the patronage system, which threatened the central power of the emperor (Webber 1989). Thus, the Chinese model set the precedent of using tests as a competitive selection device.

This purpose was repeated in Lord Macaulay's (1853) proposal to use examinations rather than patronage as a method of choosing cadets for the Indian civil service, in a similar system established in the nineteenth century for the Prussian civil service, and in current admission procedures for British and American universities. The Chinese test and the Indian civil service examinations were long and multifaceted, with many parts.

Using a test to provide information on the quality of the "product" of an education system also has a reasonably long history. A visitor to the academy in Sura, Babylonia, in the early tenth century reported that the Gaon (the head of the Yeshiva) examined the students every year and

reduced the stipend of those who had been "lazy or negligent" in their studies (Brody 1998).

Another medieval example, cited by Madaus (1990), was Treviso, an Italian town where the schoolmaster's annual salary was set on the basis of his pupils' performance on a test given at the end of the year.

In the United States, Harvard college had a statute from 1650 requiring that each year students were to be publicly examined in their knowledge of Latin, Greek, Hebrew, rhetoric, logic, and physics (Buck 1964). By 1790, professors were to conduct these examinations "in the presence of a joint Committee of the Corporation and Overseers."

Our present entanglement with examinations for national standards is the latest example of this goal of accountability. A tight system of curricular control was established in the seventeenth and eighteenth centuries in the Catholic schools, which monitored instruction through monthly examinations derived, it appears, from Jesuit observation of the Chinese example (De la Salle 1838). When church schools were secularized and nationalized in Revolutionary France, the examination system was perfected by Napoleon as a method of controlling a centralized education system (Anderson 1975). In England, at the end of the nineteenth century, elementary school examinations administered by visiting inspectors were used to justify the expense of a public education system.

The origin of a third purpose of testing—that of certifying that an individual has achieved a specific level of technical or professional skill—may date back to the first time that a parent observed a child's successful performance of some skill and decided that the child was ready to carry on without supervision. It was formalized in tests given at the end of apprenticeships, was introduced by Samuel Pepys for promotion to the rank of lieutenant in the Royal Navy at the end of the 17th century (Tomalin 2003) and has been extended to the many areas in which public certification of skill is considered socially or legally desirable.

A fourth purpose of testing, again one with a fairly natural informal beginning, is prediction or prognosis of the probable results of training. The obvious examples for those of us in the foreign language field are the prognosis tests developed in the United States to try to decide which students should be kept out of foreign-language classes because they might

increase failure rates (Henmon et al. 1929), and the aptitude tests developed in the 1950s to decide which candidates to admit to expensive government language training programs (Carroll 1962).

A fifth purpose of testing is an integral part of all good teaching: the process by which teacher and learner check the need for and progress of instruction. Whether in the form of a diagnostic test to decide what needs to be taught (Spolsky 1981, 1992) or an achievement test to check the success of teaching and learning, pedagogical tests are ideally low-stakes events that threaten none of the participants. The low-stakes nature of such tests gives the test-designer the greatest freedom to experiment. Of course, the effectiveness of this kind of testing is sometimes threatened or destroyed by adding an extra purpose.

These purposes and other matters in language assessment are well reviewed in three recent encyclopedias concerned with language teaching (appendix 1).

Measurement, fairness, and the deification of reliability

The first four purposes discussed above share a common feature that raises, or should raise, our concern for fairness. All of them assert and depend on a *power relationship* (Foucault 1975) between tester (or test user) and test taker, with the former usually given full control of the form of the test and the criteria for interpreting the test results. In language testing, language teachers and others have become increasingly concerned with the power and impact of tests, especially when used for gatekeeping purposes (Hamp-Lyons 1997; Shohamy 1992, 1999).

It was the concern for fairness in high-stakes testing that led to the development of the psychometric enterprise. When tests were used to classify candidates, there was reason to worry that manipulation of the system might favor certain candidates over others. This remained a somewhat nebulous issue until the quantification of test results—the award of numerical marks rather than of a pass—encouraged formal analysis.

Statistician Francis Y. Edgeworth (1888, 1890) argued that examinations—then widely regarded as only a rough test of merit—could be made more precise by applying the theory of errors, a branch of probability theory. Physicists, he noted, had already demonstrated the existence of error in the measurement of time, distance, and weight. A series of

measurements had been shown to deviate regularly from the correct measure, forming a normal curve—like a gendarme's hat, as one French mathematician had described it. The same phenomenon should be found in the marks given to Latin prose by different examiners. Variation might result from the health of the examinee or the selection of questions. It would also inevitably result from the limit in the degree of quality that any human being could perceive, which Edgeworth estimated to be about five percent.

The mean judgment of several competent examiners would provide the *true* score, something impossible to measure physically. In two papers (1888, 1890), Edgeworth analyzed the marking of several competitive examinations and calculated the risk in setting cut-off points. The security level, he believed, should be four times the average discrepancy between examiners. He then asserted what he termed "the unavoidable uncertainty" of tests.

Many ignored the challenge that Edgeworth posed to the testing enterprise. After a brief flirtation with "objective testing" embodying Edgeworth's principles in the early twentieth century, British examination boards managed to ignore reliability until quite recently.

The American public, by contrast—influenced by glowing but inaccurate reports (Yerkes 1921) of the usefulness of the Army Alpha tests used briefly and with little effect during the First World War—was quickly convinced that objective tests were reliable and accurate methods of measuring human mental abilities. A testing industry started to develop in the 1920s. Buttressed from criticism by esoteric psychometric techniques, it soon persuaded the public that examinations were not just powerful but could be fair.

Of all fields, language testing was perhaps most resistant to the claims of objectivity, for the techniques it demanded—breaking down language ability into the tiny discrete points that a multiple-choice or true-false test required—seemed alien to the kind of integrative performances normal in language use. Furthermore, providing the large number of judges of integrative performance needed to achieve reliability was exorbitantly expensive for a relatively low-valued skill.

The first half of the 20th century: The pursuit of objectivity

Although there had been earlier attempts at objective language testing, it was the committee appointed in 1913 by the Association of Modern Language Teachers of the Middle States and Maryland that first attempted to tackle the objective psychological testing of spoken language, a critical concern in the philosophy of the Direct Method that the association embraced. The test that the committee produced in 1914 included a dictation, written answers to questions read aloud, and the written reproduction of a passage read by the examiner (Committee on Resolutions and Investigations 1917). The assumption was that only a candidate with training in the spoken language could handle these written tasks.

The 1928 Modern Foreign Language Study (Coleman 1929) faced much the same problem. The team headed by Henmon (Henmon 1929) produced what they called the Alpha tests, which included discrete items in vocabulary and grammar as well as more integrative reading and writing tasks, the latter to be scored by comparison with 16 graded sample essays. They had no luck, however, in finding a satisfactory method of testing spoken language. They also made a start on prognostication (Henmon and others 1929) but, despite their best endeavors, failed to come up with a test that could help teachers reduce the "mortality rate" of students allowed into their courses (Cheydleur 1932).

In a long and regrettably unpublished memorandum written nearly 50 years ago, psychologist John Carroll (1954) sketched the history of foreign language testing, the current state of the art, and the areas in which research was needed. Paper-and-pencil tests of vocabulary, reading, and grammar were "highly perfected," but tests of oral and aural ability were underdeveloped. Carroll also raised interesting questions about the existing tests.

Over the next decade, as various improvements were made (Lado 1961), Carroll himself helped fill three important gaps in language tests. In the early 1950s, he tackled the problem of language aptitude, his goal being to provide an effective screening device for intensive language programs like those offered by the Army Language School and the Foreign Service Institute. The battery of tests that he developed was ready by 1955 and published commercially in 1957 (Carroll and Sapon 1955, 1957).

During the same period, Carroll was advising the Foreign Service Institute, where Claudia Wilds and her colleagues (1975) were developing an instrument to assess the language competence of State Department employees (Rice 1959; Sollenberger 1978). Although there is no record of the advice that he gave, his likely contribution may be inferred from his 1954 memorandum. In that paper, he discussed scaling and urged the development of "quasi-absolute" scales, which had been proposed but never used during the war (Kaulfers 1944; Sandri and Kaulfers 1945). He also recommended a "controlled conversation." The Foreign Service Institute's oral interview and absolute proficiency scales developed over the next few years became the core and model for many subsequent efforts at assessing spoken language ability (North 1992).

As the decade ended, Carroll (1961) set out what he believed to be the fundamentals of language proficiency testing. After acknowledging the value of Lado's description of discrete-item testing, Carroll added a plea for integrative tests. Many scholars believe this paper marked the true beginning of the language testing field. Research over the past half-century can be seen as an effort to meet the challenges that he presented (Bachman 1990).

From the point of view of language testing theory, the key question remains the nature of what we call language proficiency (the more reasonable term "language competence" having been preempted by Chomsky for an unrelated purpose) and whether it is unitary or divisible into distinct components. That issue remains unresolved.

Carroll's own work in the study of human cognitive abilities (Carroll 1993) led him to believe that foreign-language ability had distinct and measurable components, but this belief has not been established empirically. A quarter of a century after John Oller (1976) presented arguments for the existence of unitary language competence, he returned to that claim and to the claim that the same competence underlies performance on nonverbal tests (Oller 2000). The remarkably high correlations to be found between different kinds of language tests, which Carroll in 1954 tentatively attributed to the importance of vocabulary, continues to confound those who seek distinct abilities.

Can language proficiency be measured—or only judged? The answer depends, of course, on how you define it. Some aspects of proficiency are clearly measurable. There are ways to estimate how many words a language learner knows, or to estimate the percentage of morphological errors that he or she makes. Other aspects, however, require a judgment, as when we assess the quality or freshness or success of a piece of writing or a conversation. Olympic events offer an analogy. In many, the winner is determined by a measurement—the runner who is fastest, the jumper who jumps longest or highest, the thrower who hurls an object farthest, the team that scores the most goals. In others, the winner is the athlete who receives the highest rating from judges—as in boxing, diving, and equestrian events. So it is with language assessment: some aspects can be measured, but others need to be judged.

There are, as Edgeworth noted, problems in determining the true result even with measurements. Before races were timed electronically, an Olympic event would be timed by several judges, each with a stopwatch, and the "correct" result was the average of the times they recorded. Similarly, we determine the fair result of a judgment by averaging the scores awarded by a number of qualified judges. Much of the criticism of traditional essay examinations was based on evidence that different judges make different judgments, and even that the same judge makes different judgments on different occasions. To obtain a fair result, then, one needed to use more than one judge, increasing the cost of the assessment procedure.

The higher costs of multiple judges favored objective multiple-choice tests over open-ended instruments. The decision not to include a writing test in the original version of the Test of English as a Foreign Language (TOEFL) appears to have been strictly economic. Interestingly, the person at the Educational Testing Service who argued successfully against inclusion was at the same time conducting research that led to the restoration of a writing unit to the College Board's English test (Spolsky 1995).

Finding complexity

Leaving aside technical developments and increasingly complex statistical models that have helped in ascertaining the usefulness of test items, the most important development in language testing over the past half-century has probably been the recognition of its social and political context.

Contextualizing language testing—multidimensionality rampant

While John Carroll was working with linguists in the 1950s to produce the body of knowledge that became psycholinguistics (Carroll 1951), other scholars were fixing language in its social context. It was a sociolinguist with training in educational psychology, who first proposed adding the social dimension to language testing (Cooper 1968). That dimension received increasing emphasis with the spread of the "communicative approach" to language learning, which emerged after structural linguistics and behavioral psychology failed to solve the problems of language teaching (Canale and Swain 1980).

Since the 1980s, the language community has realized that tests must assess performance of authentic language functions, but those terms have yet to be satisfactorily defined and placed in an accepted theoretical model. Models have been proposed, but they turned out to be programmatic and heuristic rather than rigorous and testable.

After the discovery of context, the second major breakthrough in language testing was the recognition of the political power of tests (Shohamy 2001) and the renewed interest in the impact of examinations on the teaching process (Wall and Alderson 1993). A century earlier, Henry Latham (1877) had characterized examinations as "an encroaching power" that was blurring distinctions between liberal and technical education and narrowing the range of learning by forcing students to cram for examinations. Teaching in England, he complained, was becoming subordinate to testing, just as it was in France.

One of the most encouraging developments that followed this realization was renewed concern for ethical considerations in language testing. How, we are now expected to ask, will the tests we develop be used? How will the results be interpreted? What effects will they have on the instructional process? What effects will they have on the future of those who take them? The emphasis is now moving to test use (Bachman 2004).

From measurement to assessment

The testing profession made an unfortunate choice a century ago, when it set out to minimize the "unavoidable uncertainty" of tests rather than trying to mitigate the effects of that uncertainty and to control tests'

"encroaching power." Having made that choice, testing professionals focused on building new and better tests rather than on the demands set by the testing purpose.

Let me give a simple example. In the 40 years that TOEFL has been a major moneymaker for the Educational Testing Service, the uses of test results have been surveyed just once. That survey showed that virtually none of the test users had done the kind of study that is necessary for valid interpretation of results. Hardcastle (2000) reminds us that "there is not a significantly discernible relationship between language proficiency as defined by the TOEFL test and subsequent measures of overall academic achievement."

The Foreign Service Institute's Oral Proficiency Interview, by contrast, is an excellent example of a purpose-driven test. The deputy undersecretary of state ordered the Institute to develop a system to assess the level of proficiency of Foreign Service officers. The test used a scale that described transparently the way a candidate could be expected to function using the language. Because those who were tested were colleagues of (and usually senior to) those conducting the tests, the procedures and results had to be fair and easily justified (Sollenberger 1978).

Too often, the purpose of a test is forgotten or disguised. TOEFL, like its two predecessors, was developed to plug a loophole in the 1924 Immigration Act, which was intended to cut down on immigration from areas other than northern Europe. The Act permitted special visas for foreigners whose only purpose was study at a school or college in the United States. Three times—in 1930, 1947, and 1961—the government asked for a test that would filter out unqualified applicants. In the same way that the Scholastic Aptitude Test was sold as a fair and efficient method of controlling access to higher education for native-born Americans, so TOEFL was sold as a fair and efficient method of screening foreigners.

A wise student of mine once remarked that it is much easier to develop a new test than to explain what any existing test really measures. The absence of a good theory does not preclude practice: bumblebees can fly even though they (and even we) don't know how. The issues raised by language testing researchers will no doubt keep us busy for a long time but will not prevent us from designing, administering, and interpreting the results of tests. In this situation, the most critical issue is to appreciate that

even the most carefully designed test produces uncertain results; therefore, we need to know how to balance the need for more certainty with the cost of the results to all concerned (Elder et al. 2001).

The work with language-aptitude testing provides a good example. It became clear early on that the best way to determine whether a candidate would benefit from prolonged and intensive language instruction was to see what happened in a one-week pilot experience. The aim of the short aptitude test that Carroll developed was to filter out candidates after the pilot session and so save money and frustration. Used for this limited purpose, Carroll's test met its goal—but there was no justification for expecting it to control admission to any kind of language course.

Unfortunately, many high-stakes tests and examinations cannot guarantee the validity or comparability of their results, which is why it is so important to develop professional ethical guidelines and a code of practice for language testers (Davies 1997).

Just what is language proficiency?

The question of what it means to know a language is certainly not a new one. It is clearly related to, but different from, the question in linguistic theory of what a language is. The discrete-item approach to language testing seemed to assume that if you knew the phonology and grammar and lexicon of a language, all you needed to do to build a test was to compile an appropriate sample of these items. Communicative testing turned that approach around: one now needed to know all of the situations in which language might be used.

A functional approach to testing is likely to suit more testing purposes than a structural one. Starting from the top rather than the bottom, such an approach might first list, in the proficiency guideline format, the kinds of functions that a learner might reasonably be expected to perform and then design specific tasks that represent those functions.

Let me give an example by describing the approach that a group of us took some years ago to develop a practical literacy test for soldiers. The instructional approach at the time was structural, and one teacher told us that the students, in the middle of an intensive course, could not read a certain sign because they were only halfway through the alphabet. We

designed a test consisting of a series of literacy tasks that might be expected of a soldier. The first, as I recall, was to recognize the individual letters labeling the safety lever on a rifle. Next was recognizing various signs that might be found on an army base. In designing this test, our assumption was that its purpose was not to test literacy in the abstract, but to assess the performance of a representative group of relevant tasks.

Whenever one attempts to describe exhaustively an individual's language proficiency, it quickly becomes clear how complex and demanding the task is. The reason is not just the complexity of language, but the fact that an individual's knowledge of language is dynamic rather than static, changing from time to time, from situation to situation and from interlocutor to interlocutor. We all notice how our own foreign language performance varies, fading as the day goes on and seeming to pick up during the cocktail hour. We notice also that some of us can manage formal communication much better than small talk, whereas others can handle all the social graces but choke up when presenting a reasoned argument. We know that there are some people we feel comfortable talking to, and others whose disdain for our accent or grammar quickly freezes our fluency. The more situations in which we can observe a learner in action, the more we can learn about his or her proficiency. To expect to reduce this complexity to a single score or to one point on a one-dimensional scale is folly.

Psychometric methods of analyzing tests assume unidimensionality. Some 70 years ago, the distinguished psychologist, Edward Thorndike, outlined his ambition to construct a perfectly scaled language test, such that any candidate who answered any one question could safely be assumed to be able to answer all the previous questions (Monroe 1939). His audience quickly tried to put him straight, one of them describing his daughter's proficiency in German. Having lived with him in Germany for a year, she knew much more German than her classmates in a school in England but probably would not know many of the items that had been included in the school curriculum. Thorndike never developed his language test.

Embracing multidimensionality

The European language portfolio developed by the Council of Europe (2001) best exemplifies the purpose-driven assessment approach. Rather than attempting to develop a single testing instrument with a single score, the result of which would be used for all purposes, the portfolio is a method

of recording the evidence needed to make an assessment of the candidate's probable competence in using a language for various purposes. It records not just scores and results from language tests the candidate has taken, but also descriptions and examples of his or her actual use of the language.

The European Language Portfolio is the assessment component of the Common European Framework (Council of Europe 2001), developed as part of the Council of Europe's project, "Language learning for European citizenship" (Trim 1997). The sweeping and ambitious framework describes what language users need to do to communicate in a situation, identifies texts that convey messages, details the underlying competences of the user that permit communication, and describes the strategies used to apply these competences. It also surveys alternative approaches to language learning and teaching, offers a set of proficiency scales, and discusses curricular designs to achieve various kinds of plurilingual competence.

Intended to bring about European cooperation in foreign-language teaching, the Common European Framework is an intimidating document, looking more bureaucratic than scholarly. There are no footnotes, only a short list of further readings, and no supporting evidence or data. Users are "invited" to derive practical lessons from a catholicity of approaches. Dozens of theses could—and probably will—be written to unpack concise maxims like, "The external conditions under which communication occurs impose various constraints on the user/learner and his/her interlocutors," or, "The output of the process of language production is a text which once it is uttered or written becomes an artifact carried by a particular medium and independent of its producer." No one who has read this work carefully can imagine an assessment model that will produce a single measurement scale, yielding a single score or grade or mark that would contain the complexity involved in assessing and describing plurilingual competence.

The term "plurilingual" has come to serve in the Framework as a label for the acceptance of complexity. "Bilingual," by contrast, has misled by its suggestion of a person able to function equally in two languages. Plurilingual competence implies not one or more languages added to the native language, but a competence that can draw on more than one language for communication. Competence in each language varies and is

uneven and dynamic. So, too, must assessment models be varied and dynamic.

The European Language Portfolio, now in the advanced design stage, consists of three parts. Part I records formal qualifications "in an internationally transparent model," drawing on an agreed proficiency scale that reinterprets national scales. Part II is a language biography, an organized account of language learning and use experiences, and a self-assessment. Part III is a dossier in which a learner can present examples of his own work using the language.

Very important principles underlie this elaborate model. First, its user orientation: the language learner, closely involved in presenting his or her own competence, is thus encouraged to continue developing it. Second, the model assumes that plurilinguals have complex patterns of varied competence in different domains. It avoids the trap of assuming a one-dimensional scale and setting out to rank all students on it, the trap that has ensnared most high-stakes testing. But it must be noted that criticisms are starting to appear of the lack of empirical validation and the trend to rigidity of this potentially open model (Fulcher 2004).

Given its financial and institutional robustness, the psychometric industry will no doubt continue to try to reduce "unavoidable uncertainty" and to develop better measures of identifiable and relevant competences. For my part, I will put more stock in approaches such which seek ways to live with uncertainty and to develop ethically based, use-oriented methods of assessing language competence.

Appendix 1 Suggested readings

Three recent encyclopedias concerned with language teaching provide good coverage of the field of language assessment at the beginning of the millennium.

One of the eight volumes of the Kluwer *Encyclopedia of Language and Education* is devoted to language testing and assessment (Corson 1997). It contains 29 articles ranging from testing reading in the mother tongue to ethics in language testing. Its summary of trends makes the point that a volume produced a decade earlier would have concentrated more on "receptive and integrative methods of assessment" and on developments in psychometrics. Now, productive skills are more important, and there is

more emphasis on how well various traits are measured. A new edition is being prepared and should appear in 2006.

The Elsevier *Concise Encyclopedia of Educational Linguistics* (Spolsky 1999) expands on the rather condensed treatment given language testing in the large edition by adding five new articles, three of which deal with alternative assessment, the impact of language testing, and the uses of language tests.

The Routledge *Encyclopedia of Language Teaching and Learning* (Byram 2000) includes assessment and testing as an overview article and—after sections dealing with alternative assessment, impact, and ethics—concludes that "the competing requirements of test validity and financial practicality will maintain the distinction between tests which can be administered reliably to large numbers of students, and more holistic tests which can potentially reveal all aspects of the candidates' language proficiency." It also includes 14 short articles on testing topics.

The growing importance of the field is also shown by the number of new books dealing with it. Cambridge University Press now has two separate series of books on language testing— eight books in the Language Assessment series (three promised for 2005) and fourteen in the Studies in Language Testing Series —as well as half a dozen books on assessment in other series. These books report on current technical research and summarize the assessment of listening, writing, reading, and vocabulary for special purposes.

A recent introduction to the field of language testing is McNamara (2000). Brief and up-to-date, it sets out to show that language testing is not "an arcane and difficult field, and politically incorrect to boot." After surveying the field and treating the social character of language tests, the volume concludes with sections on the use of computers, the problem of finding a cheap way to assess speaking ability, and the dilemma of assuming who is responsible for breakdowns in real-life communications. Other new books include (Weir 2005), (Brown 2003) and (Puerschel & Raatz 2001).

Over its first 80 years of publication, the *Modern Language Journal* published about 150 papers on language testing, fewer than 2 per volume (Spolsky 2000). These articles trace the historical growth of language testing and the profession's reluctant recognition of the fact that language

tests both drive and reflect language teaching. Testing oral proficiency has been an ongoing theme from the earliest volumes through to the *ACTFL Proficiency Guidelines*. For those to whom the spoken language is a key part of the curriculum, finding an efficient way of assessing ability in this area has been a continuing challenge. Other important themes recur regularly: cloze tests, proficiency guidelines, prognosis and aptitude testing, and test use.

References

Anderson, R. D. (1975). *Education in France 1848–1870*. Oxford: Clarendon Press.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, Lyle G. (2004). Building and supporting a case for test use. Language Assessment Quarterly: An International Journal, 2(1), 1-34.

Brody, R. (1998). *The Geonim of Babylonia and the shaping of medieval Jewry*. New Haven: Yale University Press.

Brown, H. Douglas. (2003). *Language assessment: principles and classroom practices*. New York: Longman.

Buck, P. H. (1964). *Examinations: a retrospective view at Harvard*. In L. Bramson (Ed.), *Examining at Harvard College*. Committee on Educational Policy, Harvard University.

Byram, M. (Ed.) (2000). *Routledge Encyclopedia of language teaching and learning*. London: Routledge.

Canale, Michael, and Merrill Swain (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics* 1, 1–47.

Carroll, John B. (1951). Report and recommendations of the interdisciplinary summer seminar in psychology and linguistics at Cornell University, June 18–August 10, 1951. Ithaca, NY.

Carroll, John B. (1954). Notes on the measurement of achievement in foreign languages. Unpublished manuscript.

Carroll, John B. (1961). Fundamental considerations in testing for English language proficiency of foreign students. In *Testing the English proficiency of foreign students*. Center for Applied Linguistics, Washington, DC.

Carroll, John B. (1962). The prediction of success in intensive foreign language training. In R. Glaser (Ed.), *Training research and education* (pp. 87-136). Pittsburgh: The University of Pittsburgh Press.

Carroll, John B. (1993). *Human cognitive abilities: a survey of factor-analytic studies*. Cambridge: Cambridge University Press.

Carroll, John B., and Stanley M. Sapon (1955). Psi Lambda Foreign Language Aptitude Battery. Laboratory for Research in Instruction, Graduate School of Education, Harvard University.

Carroll, John B., and Stanley M. Sapon (1957). *Modern Language Aptitude Test*. New York: Psychological Corporation.

Cheydleur, F. D. (1932). Mortality of modern languages students: Its causes and prevention. *Modern Language Journal* 17, 104–136.

Coleman, A. (1929). *The teaching of modern foreign languages in the United States*. New York: Macmillan.

Committee on Resolutions and Investigations (1917). Report of committee on resolutions and investigations appointed by the Association of Modern Language Teachers of the Middle States and Maryland. *Modern Language Journal* 1, 250–261.

Cooper, Robert L. (1968). An elaborated language testing model. *Language Learning* 57–72.

Corson, David. (Ed.) (1997). *Encyclopedia of Language and Education*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

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

Davies, Alan. (1997). Introduction: the limits of ethics in language testing. *Language Testing, 14*(3), 235-241.

De la Salle, J.-B. (1838). *Conduite des ecoles chrétiennes*. Paris: J. Moronval.

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.

Elder, Catherine, Brown, Annie, Grove, Elisabeth, Hill, Kathryn, Iwashita, Noriko, Lumley, Tom, et al. (Eds.). (2001). *Experimenting with uncertainty: Essays in honour of Alan Davies*. Cambridge: Cambridge University Press.

Foucault, M. (1975). *Surveiller et punir: Naissance de la prison*. Paris: Gallimard.

Fulcher, Glenn. (2004, 17 May). Are Europe's tests being built on an 'unsafe' framework? *Education Guardian*.

Hamp-Lyons, L. (1997). Washback, impact and validity: Ethical concerns. *Language Testing* 14, 295–303.

Hardcastle, P. (2000). How not to test language. *Language Testing Update* 28, 18–24.

Henmon, V. A. C., J. E. Bohan, C. C. Brigham, L. T. Hopkins, G. A. Rice, P. M. Symonds, J. W. Todd, and R. J. Van Tassel (Eds.) (1929). *Prognosis tests in the modern foreign languages: Reports prepared for the Modern Foreign Language Study and the Canadian Committee on Modern Languages*, vol. 16. New York: Macmillan.

Henmon, V. A. C. (1929). *Achievement tests in the modern foreign languages, prepared for the Modern foreign language study and the Canadian committee on modern languages*. New York: Macmillan.

Kaulfers, W. V. (1944). Wartime development in modern-language achievement testing. *Modern Language Journal* 28, 136–150.

Lado, Robert. (1961). *Language testing: The construction and use of foreign language tests—A teacher's book*. New York: McGraw Hill.

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

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

Madaus, G. P. (1990). Testing as a social technology. Presented as the first annual Boise lecture on education and public policy, Boston College.

McNamara, T. (2000). *Language testing*. Oxford: Oxford University Press.

Monroe, P. (Ed.) (1939). Conference on examinations under the auspices of the Carnegie Corporation, the Carnegie Foundation, the International Institute of Teachers College, Columbia University, held at the Hôtel Royal, Dinard, France, September 16–19, 1938. Teachers College, Columbia University.

North, Brian. (1992). Options for scales of proficiency for a European Language Framework. Occasional paper series, National Foreign Language Center, Washington, DC.

Oller, John W., Jr. (1976). Evidence of a general language proficiency factor: An expectancy grammar. *Die Neuen Sprachen* 76, 165–174.

Oller, John W., Jr., K. Kim, and Y. Choe (2000). Testing verbal (language) and nonverbal abilities in language minorities: A socio-historical problem in historical perspective. *Language Testing* 17, 341–360.

Puerschel, Heiner, & Raatz, Ulrich (Eds.). (2001). *Tests and translation: Papers in memory of Christine Klein-Braley*. Bochum, Germany: AKS-Verlag Bochum

Rice, F. (1959). The Foreign Service Institute tests language proficiency. *Linguistic Reporter* 1, pages 2, 4.

Sandri, L., and W. V. Kaulfers (1945). An oral-fluency rating scale in Italian. *Italica* 22, 133–144.

Shohamy, Elana. (1992). The power of tests: A study of the impact of language tests on teaching and learning. Paper presented at the Language Testing Research Colloquium, Vancouver, BC.

Shohamy, Elana. (1999). Language testing: Impact. In Bernard Spolsky (Ed.), *Concise encyclopedia of educational linguistics*. Amsterdam: Elsevier.

Shohamy, Elana. (2001). *The Power of tests: a critical perspective of the uses of language tests*. London: Longman.

Sollenberger, H. E. (1978). Development and current use of the FSI Oral Interview Test. In J. L. D. Clark (Ed.), *Direct testing of speaking*

proficiency: theory and application. Princeton, NJ: Educational Testing Service.

Spolsky, Bernard. (1981). The gentle art of diagnostic testing. Paper presented at the Interuniversitätsprachtestgruppe Symposium, Hasensprungmuehle, Germany, December 15.

Spolsky, Bernard. (1992). The gentle art of diagnostic testing revisited. In E. Shohamy and R. Walton (Eds.), *Language assessment for feedback and other strategies*. Washington, DC: National Foreign Language Center.

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

Spolsky, Bernard. (2000). Language testing in the *Modern Language Journal*. *Modern Language Journal* 84, 536–552.

Spolsky, Bernard. (Ed.) (1999). *Concise encyclopedia of educational linguistics*. Amsterdam: Elsevier.

Tomalin, Claire. (2003). *Samuel Pepys: the unequalled self*. London: Viking.

Trim, John. (1997). Final report of the project on language learning for European citizenship. Council of Europe, Strasbourg, France.

Wall, Dianne, and J. C. Alderson (1993). Examining washback: The Sri Lankan impact study. *Language Testing* 10, 41–69.

Webber, C. (1989). The mandarin mentality: Civil service and university admissions testing in Europe and Asia. In B. R. Gifford (Ed.), *Testing policy and the politics of opportunity allocation: The workplace and the law*. Boston: Kluwer Academic Publishers.

Weir, Cyril. (2005). *Language testing and validation: an evidence-based approach*. Basingstoke UK: Palgrave Macmillan.

Wilds, Claudia. (1975). The oral interview test. In B. Spolsky and R. L. Jones (Eds.), *Testing language proficiency*. Washington, DC: Center for Applied Linguistics.

Yerkes, R. M. (Ed.) (1921). *Psychological examining in the United States Army*. Washington: Government Printing Office.