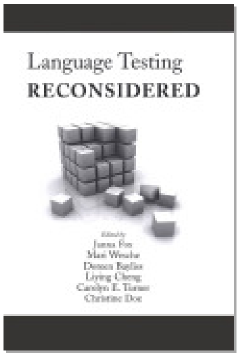
PROJECT MUSE®

Language Testing Reconsidered

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



Language Testing
RECONSIDERED

Edited by
Janna Fox
Mari Wesche
Doreen Bayliss
Liying Cheng
Carolyn E. Turner
Christine Doe

Published by University of Ottawa Press

Fox, Janna & Wesche, Mari & Bayliss, Doreen & Cheng, Liying.
Language Testing Reconsidered.
Ottawa: University of Ottawa Press, 2007.
Project MUSE., https://muse.jhu.edu/.

1 ON SECOND THOUGHTS

Bernard Spolsky
Bar-Ilan University

Abstract

An elderly student of language testing recalls some of the events in his career, attempting to relate them to current issues, and touching on such topics as overall language proficiency, the cloze and the noise test, the social responsibility of language testers, the development of industrial language testing, the danger of scales, and the value of knowing the history of one's field.

When I was awarded the rank of professor emeritus a few years back, a colleague kindly pointed out the real translation of the term: "e-" means out, and "meritus" means deserves to be. And when I learned in 2005 that the International Language Testing Association had decided to award me its lifetime achievement award, I took it as a clear confirmation that productive work was over.[1] I therefore feel no challenge to present new research here (in point of fact, the research that I have been doing since I retired has dealt not with language testing but with language policy and management) but to present some second thoughts on work that I did some years ago.

A Testing Tyro

In point of fact, I got into language testing through the backdoor: as a teacher, I became more and more suspicious of the relationship between tests and their results.[2] As a teacher-administrator responsible for a program of English for foreign students at Indiana University, my alarm was first aroused by the way in which we were using various kinds of tests. In particular, I became worried about the social consequences of using a test of English as a criterion for admission of foreign students to universities.[3] In an early paper, delivered at a conference of Administrators and Teachers of English as a Second Language (ATESL), the English teaching section of the National Association of Foreign Student Advisors: Association of Internatinal Educators (NAFSA) that preceded the Teachers of English to Speakers of Other Languages (TESOL), I raised a socio-ethical question: were we not, by using English proficiency as a criterion for admission of foreign students to American universities, limiting study in the United States to the children of parents well enough established financially or politically to make sure that they could go to the small number of good schools in their country which taught English reasonably well? At the time, in the late 1960s, it was generally held by English teaching professionals that English was for the elite: I heard many United States Information Agency (USIA) and British Council experts decrying the growing tendency to meet the burgeoning demand for English throughout the world by trying to teach it to everyone.[4] This experience confirmed my first realization, as a high school teacher in a largely Maori area in New Zealand, that teaching takes place in a social context and with constant social implications. It was shortly after that that I heard Robert Cooper's pioneering introduction of sociolinguistic criteria into language testing.[5]

But before I could fit that into my own work on language testing, I was attracted into the increasingly popular pursuit of the will-o-the-wisp of overall proficiency. During the time that I was at Indiana University, I became friendly with Bengt Sigurd, a recent graduate from the University of Lund who held a visiting lectureship at Bloomington.[6] In a conversation that we had after an otherwise unmemorable lecture, we wondered whether it might not be possible to test the overall proficiency of a second language learner by adding noise to a taped text. Thus was born the noise test,[7] and subsequent explorations of the relevance of reduced redundancy as a language testing tool.

In an effort to answer what has remained my fundamental research question ("What does it mean to know a language?"), we made use of the concept of redundancy developed as part of the statistical theory of communication (Shannon and Weaver, 1963). Three techniques, I argued in a paper that appendicitis prevented me reading at the Second International Congress of Applied Linguistics at Cambridge, England, in 1969,[8] had been developed that took advantage of this principle. One was the cloze, which had been proposed by Taylor (1953) as a way of determining the readability of a text, considered but rejected as a language testing technique by Carroll (Carroll, Carton and Wilds, 1959) in an unpublished paper,[9] and beginning to be used for language testing by Holtzman (1967) and others but not yet fully explored as it was to be by Oller (1972) and his students. The second was an intriguing attempt at a standardized cloze test called clozentropy (Darnell, 1968, 1970).[10] The third was the noise test itself (Spolsky, Sigurd, Sato, Walker, and Aterburn, 1968).

Each of these techniques had interesting continuing development. Oller's work with the cloze is well known,[11] and some of the doubts raised with it led to the subsequent development of the C-test (Klein-Braley, 1997). Douglas (1978) among others like Lowry and Marr (1975) kept up the interest in clozentropy. Gaies, Gradman and Spolsky (1977) followed up with the noise test. My own comfort with it faded when evidence emerged that some test-takers had excess of anxiety with the task of dictation with noise added.[12]

Finding Historical Context

Thinking back to those days, I recollect with sincere thankfulness my interactions with colleagues and students. My first important research was initiated in conversations with a visiting colleague and carried out with the collaboration of students. The first major language testing conference that I attended was at the University of Michigan, organized by Jack Upshur; other meetings followed regularly and finally were transformed into LTRC. What I later learned is that these meetings could be traced further back (see Spolsky, 1995, pp. 158–159); the first major language testing meeting was the session organized by John Carroll at the 1953 Georgetown Round Table (Hill, 1953), about which I had read while learning the historical context of the field.

I only learned about that meeting when I started on my research for the book that appeared 10 years ago (Spolsky, 1995). This originally was intended to be a study of the development of the Test of English as a Foreign Language (TOEFL). As I started to read the papers written for the initial planning conference (Center for Applied Linguistics, 1961) and especially the paper considered by many to be the first major work in language testing (Carroll, 1961), it became clear to me that I really did not know the historical background.

Coming into language testing in the mid-1960s, I had a fairly elementary view of the history of the field. With the daring of ignorance, I even had the nerve to put forward that view in a keynote address at the Association internationale de linguistique appliquée (AILA[13]) Congress in Stuttgart (Spolsky, 1977). I proposed that language testing had gone through three stages: an undefined stage of traditional testing, a stage during which structuralist linguistics and psychometrics had agree on a method of assessing knowledge of individual linguistic items, and the current period during which integrative testing (the term came from Carroll's paper) and sociolinguistics were leading to a new approach. The oversimplification appears to have been appealing and is still widely cited. But when in the 1990s I went back to read those earlier papers, I quickly realized how seriously I was suffering by not having learnt the history of my field.

Take a simple example. Many of us believed that Carroll's 1966 paper condemned Lado for not treating integrative testing techniques; read more closely, one sees that in fact he praised Lado (1961) for having adequately covered the field of item testing, making it appropriate to him to describe the additional value of integrative testing. I tried to make up for my mistaken assumption in Spolsky (1996) and in organizing the LT+25 conference, which celebrated the 25th anniversary of the appearance of Carroll's and Lado's pioneering works (Carroll, 1986).

So I ventured into history. For the first several months of a sabbatical, supported in part by a Mellon Fellowship from the Institute of Advanced Studies of the National Foreign Language Center (NFLC) at Johns Hopkins University,[14] sitting in a building in Washington, DC, just a block from where the 1961 meeting I was studying had taken place, or working at the Library of Congress or at the Georgetown University Library,[15] I sat learning what I could about the development of language testing in the years before I came into the field. The review took up half the book and gave me, I believe, a much clearer understanding of the development of our field.

This volume, with its interest in the past, is surely a place to complain about our lack of historical sense. Since my own excursion into the field, I have continually talked about history, finding in the 2000-year-old Chinese examination tradition,[16] or in the mediaeval northern Italian testing technique,[17] or in the late 17th century Royal Navy examination for promotion to the rank of lieutenant,[18] or in the mammoth examination introduced for the Indian Civil Service in the middle of the 19th century,[19] inspiration and explanation for current approaches to language testing. There are some scholars now, I am happy to say, who appreciate the historical approach. When one of them told me that he was teaching a course on the history of language testing, I sent him a manuscript copy of the first part of my book: he thanked me but noted that he planned to start his course just where I left off.

Of course, in general, we tend to cite only the most recent work in our field. For example, a recent journal issue devoted entirely to the ethics of language testing manages to cite Spolsky (1997) but ignores completely Spolsky (1981, 1984) in which I first use the term. And of course no one mentioned Spolsky (1967) which raised ethical questions about language testing. In the same way, discussions of reliability no longer refer back to Edgeworth (1888, 1890), or of validity of essay-marking to Sir Phillip Hartog (Hartog and Rhodes, 1935, 1936), or of the problem of scaling to the elderly Thorndike's dream of an absolute scale for language proficiency (Monroe, 1939), nor do our criticalists cite the impassioned attacks on the "encroaching power" of examinations expressed by Henry Latham (1877).

I suspect that it is our lack of historical sense and knowledge that condemns us to continually rediscovering the wheel. The current enthusiasm for testing is not new: in 1882, the Gilbert and Sullivan opera included this verse:

> Peers shall teem in Christendom,
> And a Duke's and exulted station
> Be attainable by com-
> Petitive examination.[20]

But competitive examinations by their very nature were limited to a small part of the population: like the Indian Civil Service examination or the Cambridge Tripos on which it was modeled, they aimed to select the very top candidates among a chosen elite. Psychometric theory made clear, as Edgeworth noted, that it was easier to make decisions at the extremes: a candidate who scored in the top 5% was clearly better than others. But at the centre of the curve, where most candidates are usually bunched, the standard error makes it very difficult to set precise boundaries.[21] This of course is just what happened when examinations were moved from the task of selection of the very best to the task of monitoring of the masses. This happened, it will be recalled, in the latter part of the 19th century in England, when the enthusiasm expressed by Gilbert and Sullivan was applied to developing tests to evaluate the effectiveness of mass elementary education. The door was thus opened both for efforts to establish standardized testing on the one hand and the development of the testing industry on the other.

Relations with Industry

Looking back, I am not sure how much I appreciated these changes before I made my excursions into history. As with many others in the field, most of my testing research was involved in developing small local tests. I certainly never assumed that the noise test would have commercial application, or that the functional tests that I worked on with my students (Spolsky, Murphy, Holm and Ferrel, 1971) would have other than local application.[22] But I was able to observe and to start worrying about the effect of industrialized testing even before I wrote my study. This was in 1967, when I became a consultant to the Educational Testing Service and the College Entrance Examination Board (they jointly owned TOEFL at that time) as member and later chair of the Committee of Examiners for the Test of English as a Foreign Language. During those three years, in short two- or three-day visits to the luxurious campus outside Princeton, I had a chance to learn the strengths and weaknesses of a large, powerful testing organization. Leaving aside the comforts (we were accommodated in the Princeton Inn, far beyond the Holiday Inn an assistant professor could normally afford), there was the opportunity of concentrated discussion about testing, not just with fellow university-based language testers but also with members of the senior research staff at Educational Testing Service (ETS).

I recall vividly several such discussions with William Angoff.[23] In one memorable icy car journey from the College Board offices in New York to Princeton, we talked about the strange phenomenon of the high correlation between the various parts of all foreign language tests: the only exception he had noted was Latin. This encouraged my continued search for overall language proficiency.

The weaknesses of industrial testing became obvious in our discussions with the test editing and production staff: they were constrained by the strict demands of the machines that controlled and evaluated their work. One such demand, as I later learned, was that every new test must be perfectly calibrated with all previous tests, a perfectly reasonable seeming requirement for a public standardized test, but one that meant that TOEFL was locked into the very first abnormal population (the particular students who were in large English as a Foreign Language (EFL) programs in the United States in the summer of 1962) on which it had been standardized. It also became clear how difficult it was to make changes in an industrial test — somewhat like steering a modern supertanker — and how little appreciation there was in ETS for TOEFL other than as a method of paying for interesting research in other domains.[24]

Pursuing a Chimera — Overall Proficiency

The conversation with Angoff brings me back to trying to understand what we were trying to do in those days. In those early days, some people in the field of language testing had come to it from psychometrics, but the majority of us came from applied linguistics — we certainly made sure that our doctoral students learnt more psychometrics and statistics than we had. There was a continuing and sometimes upsetting friction between the two, as we impatiently dealt with our areas of ignorance. At Princeton, interestingly enough, the local people were impressed and irritated by our psychometric sophistication, limited though it might have been. Unlike other committees of examiners, we were interested not just in content but also in testing. But within our own field, we were in the midst of the ongoing struggle to define the nature of applied linguistics and its role in dealing with real-world language-related problems.

Looking back, I see my own tension between the search for overall language proficiency on the one hand (witness the noise test and other work on redundancy) and the development of sociolinguistically relevant functional tests on the other as representing the same kind of bipolar attraction that the language sciences were struggling with. The language sciences were starting to bifurcate, with most following (or resisting) Chomsky as he set out to build a theory of linguistics that would justify it as a brain science, and another large segment following Labov (1972) and Fishman (1972) and others who investigated the social patterns of language use.[25] It is not too far-fetched to suggest that those language testers who pursued overall language proficiency were influenced by the language theorists, and those who started to develop functional authentic tests were influenced by the sociolinguistic model proposed originally by Cooper (1968) at the seminal meeting on language testing organized by Jack Upshur at the University of Michigan.

Unanswered Questions

Thirty years later, I still cannot give a short answer to the question, "What does it mean to know a language?" I have learned some of the characteristics that distinguish a native speaker from a learner, such as the ability to handle a reduction in redundancy. But I have also learned from colleagues (Davies, 1991) of the uncertainty of the concept "native speaker" and from my experience with the over-powerful Interagency Language Roundtable (ILR) scale, the problems produced by setting an educated native speaker as the highest level to be aimed at by a language learner.

What was originally called the Foreign Service Institute (FSI) absolute language proficiency scale (Jones, 1979b; Wilds, 1975) was developed, with the advice of John Carroll, to provide a method of encouraging diplomats to learn the language of the country to which they were being posted. The scale was aimed to rank the various language functions expected of a diplomat, starting with straightforward daily life in a foreign city, passing through the skills to handle simple diplomatic and consular business, and culminating in the ability to impress foreigners as speaking and writing the language better than they did themselves. In actual practice, it grew up as an evolving consensus between examiners and administrators, controlled all the time by the high status of their examinees, who were commonly higher on the pay scale than their language teacher-examiners. As time went on, the scale was adapted by other government agencies, each of which made their own local adjustments in scoring and administration without attempting to design their own valued set of functions. The problem produced is most easily demonstrated by the use of the scale by the Federal Bureau of Investigation. One of their tasks involves listening to the conversations of suspected drug-runners: thus, the ability to understand the colloquial Haitian Creole they speak or the mixture of Mandarin and Cantonese used by the Hong Kong gangs surely deserves to be at the top of their scale rather than closeness to the speech of a pedantic university literature professor. When developing its own graded framework of functional skills, the Common European Framework (Council of Europe, 2001) uses the term "native speaker" only once, when it set the goal of speaking in such a way that you can be understood by a native speaker.

The Council of Europe framework is exceptional in the exhaustiveness with which it sets out to list all the conceivable functional goals of foreign language teaching and all the known individual items that make up language knowledge. It is also refreshingly modest in its insistence that it is merely a guide to be used to develop a curriculum or test for specific purpose and context. In practice of course, it is no more validated than any other scale is (Fulcher, 2004a) and is as easily translated into rigidity.

The underlying question, as true of tests as of scales, is how to value results and translate them into interpretations. A fair criticism of the first 19th century spelling test was that it just counted the number of words correct without deciding which words are more important or more difficult. The work with modern vocabulary testing by Nation (1990) attempts to overcome this by using frequency as a criterion. In the early work that he did with scales, Thorndike (1910) developed a writing scale that consisted of examples of handwriting that had been ranked by several thousand teachers. Similar approaches were proposed for essays and other scaling, and in his own early work on oral testing, the pioneering Cambridge language tester John Roach (1945) made use of recorded samples to develop and train judges. While Thorndike himself agreed that averaging the scores of two separate judges leads to greater reliability, he also insisted that each be allowed to make a judgment based on their own criteria: a judge should be consistent, but there was no reason to expect judges to agree on a single scale. Clearly, he would disapprove of the techniques used by industrial testing concerns to see that their hired judges agree as closely as would computerized marking machines.

And of course the debate continues, in its most recent and sophisticated form in Bachman (2005). And it is just as well that it does, for I suspect that if ever we were to all agree on the nature of language proficiency and on how to measure it, we would simply build massive and powerful testing engines that would rapidly pigeonhole all our students.

Our lack of historical sense makes me wonder sometimes whether or not we are making progress. Attending our meetings, reading our papers and books, and comparing them with the work of our predecessors, I sometimes suspect that we have added techniques rather than understanding. The papers in this volume however will correct this over-pessimistic view, for they show that research in the field of language testing has been producing both new questions and new answers to some of the old ones.

In this paper, more of a memoir than I intended, I have been drawn unwittingly into some of the current debates but carefully avoided others; [26] on another occasion, granted a little more space or time,[27] I hope to speculate on future trends.

Notes

[1] As in Hamlet's reference to funeral baked meats furnishing the marriage table, this paper served a double function, being a seminar contribution and an award acceptance speech.

[2] My first testing publication in 1965 was a review of two tests.

[3] My second testing publication was a paper (Spolsky, 1967) at a conference of foreign student advisers and English as a foreign language teachers.

[4] I recall the British Council and the USIA experts in Bangkok in 1967 agreeing that the way to improve the level of Thai English was to stop trying to teach it in elementary school.

[5] Cooper (1968) was a byproduct of the major sociolinguistic study of the New Jersey Barrio (Fishman, Cooper, and Ma, 1971) that was underway at the time of the Michigan meeting.

[6] Bengt Sigurd is now emeritus too, after a distinguished career in phonetics and general linguistics at the universities of Stockholm and Lund.

[7] Spolsky, Sigurd, Sato, Walker, and Aterburn (1968) was first reported on at the Michigan meeting.

[8] The subsequent publication of the paper (Spolsky, 1971) led several people to believe that I had been there. Among other sources of confusion for language testing historians, Davies (1968) was not a symposium in the sense of meeting, but a gathering of papers intended to give a picture of the state of the field (Kunnan, 2005a, p. 39); and the people listed as Department of Agriculture in Carroll's 1953 sessions at the Georgetown Roundtable (Carroll, 1953) were the early CIA language testers.

[9] He thought it more likely to be a specific ability.

[10] Essentially, he used a computer to compare a foreign student's answers to those of a selected group of American students studying a specific field.

[11] For example, J. W. Oller, Jr. (1975, 1972).

[12] It is not unreasonable to consider anxiety a form of noise. See, for instance, Vogely (1998).

[13] AILA, the Association internationale de lingguistique appliqueée, was founded at the International Colloqium of Applied Linguistics at the University of Nancy, France. The English translation is the International Association of Applied Linguistics (also known as AILA).

[14] Directed at the time by Richard D. Lambert, NFLC has since moved to the University of Maryland. Over the years, it and its offspring, the Center for Advanced Study of Language, have provided comfortable accommodation and stimulating ideas for my regular visits to Washington.

[15] Georgetown inherited a vital collection of papers from the Center for Applied Linguistics.

[16] I felt proud when I was invited to give a lecture on language testing at the Institute for Applied Linguistics in Beijing, a visit that gave me the chance to see the hall used for the final examination in the Forbidden City. And I enjoyed learning more about the Chinese origin of testing in Liz Hamp-Lyons' plenary at LTRC 2005.

[17] I learned about Treviso and its policy of paying the schoolmaster by results from Madaus and Kellaghan (1991). It was a similar "closing of the circle" when I was asked to advise on the Treviso test being developed by the University of Venice and reported on by Geraldine Ludbrook at this LTRC.

[18] One of Samuel Pepys' contributions as a naval administrator (Tomalin, 2003).

[19] The arguments for replacing patronage with testing presented by Macaulay (1853) and others are worth reading.

[20] *Iolanthe*, W. S. Gilbert, 1882.

[21] It is good to see the reference to standard error of measurement in the proposed ILTA Code of Practice.

[22] Of course, with the burgeoning of demand, a number of our colleagues have managed to start their own testing enterprises, raising a new set of ethical problems.

[23] William H. Angoff (1919–1993) was a distinguished research scientist at ETS for more than 40 years.

[24] Spolsky (2003) contrasts this with a quite different approach that developed in the 1990s at UCLES, where the language testers in charge made sure that the profits went back into language testing research and test evaluation and improvement.

[25] Labov, of course, always argued that he was doing theoretical linguistics and that the social information was needed to handle the variation that the generative linguists ignored, and Fishman of course never claimed to be a linguist but a Yiddishist studying the sociology of language.

[26] Such as the fascinating discussion of hotel bedrooms in recent lists, or the agonizing worries of government use of tests to return asylum seekers to their inhospitable homes.

[27] For example, Saville and Kunnan (2006).