This article was downloaded by: [National Institute of Education]
On: 02 May 2014, At: 22:54
Publisher: Routledge
Informa Ltd Registered in England and Wales Registered Number: 1072954
Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK

Language Assessment Quarterly
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hlaq20

INTERVIEW: People and Events in Language Testing: A Sort of Memoir An Interview With Bernard Spolsky

Nick Saville & Antony Kunnan
Published online: 16 Nov 2009.

To cite this article: Nick Saville & Antony Kunnan (2006) INTERVIEW: People and Events in Language Testing: A Sort of Memoir An Interview With Bernard Spolsky, Language Assessment Quarterly, 3:3, 243-266, DOI: 10.1207/s15434311laq0303_3

To link to this article: http://dx.doi.org/10.1207/s15434311laq0303_3

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the "Content") contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is

expressly forbidden. Terms & Conditions of access and use can be found at
http://www.tandfonline.com/page/terms-and-conditions

Downloaded by [National Institute of Education] at 22:54 02 May 2014

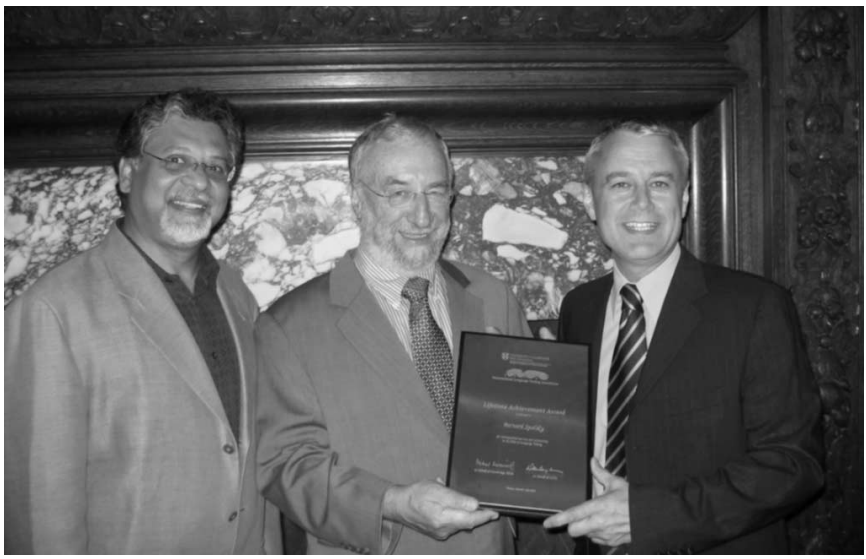
LANGUAGE ASSESSMENT QUARTERLY, *3*(3), 267–294
Copyright © 2006, Lawrence Erlbaum Associates, Inc.

Downloaded by [National Institute of Education] at 22:54 02 May 2014

INTERVIEW

People and Events in Language Testing: A Sort of Memoir An Interview With Bernard Spolsky

Nick Saville and Antony Kunnan



Bernard Spolsky after receiving the UCLES/ILTA Lifetime Achievement Award with Antony Kunnan and Nick Saville.

This interview took place at the Language Testing Research Colloquium in Ottawa, Ontario, Canada (at the Chateau Laurier Hotel on July 21, 2005), at which Professor

Correspondence should be addressed to Nick Saville, Director–Research and Validation, University of Cambridge, ESOL Examinations, 1 Hills Road, Cambridge, CB1 2EU, UK. E-mail: Saville.N@cambridgeesol.org

Bernard Spolsky was presented with the University of Cambridge Local Examinations Syndicate/International Language Testing Association Lifetime Achievement Award. The conference provided a context for the points discussed with Nick Saville and Antony Kunnan, and some references were made to people who presented papers at the conference. In recognition of the award, Professor Spolsky gave a lecture, "On Second Thought," during a symposium entitled *Rethinking Language Testing: Voices From Experience*, organized by Mari Wesche. He also made an after-dinner acceptance speech at the conference banquet when the award was made. Some of the issues discussed in the interview were presented in these two talks.

Downloaded by [National Institute of Education] at 22:54 02 May 2014

NS: Bernard, you were born in 1932. Where was that exactly? And what are your recollections of growing up there at that time?

BS: I was born in Wellington, New Zealand. My parents had been born in Britain, my father in Glasgow, and my mother in Bournemouth, and both were brought out to New Zealand by their parents. My father's parents had come to Scotland from the Ukraine by 1895 and immigrated to New Zealand in 1906, when the tobacco business was going through a slump. My mother's father, born in the London East End to a tailor who had come from Poland by 1848, immigrated to Australia in 1897 and went on to New Zealand in 1900 and, when his brother and sister died in 1904, made a trip to America the following year, where he married my grandmother. She disliked New Zealand, so he took her to England, where they stayed through the First World War and returned to New Zealand with their son and schoolgirl daughter (my mother) in 1920.

NS: When did you first develop an interest in language?

BS: Linguistically, I was brought up in an English-speaking home in an English-dominant community. My father knew but did not speak Yiddish and, not having completed high school, had learned no other language. At the same time, growing up in a moderately observant and deeply committed Jewish home—we kept kashrut, went to synagogue once a week, and both my parents spent most of their spare time in Jewish communal activities—I quickly became aware of Hebrew as a sacred language and learned to read prayers. After my bar mitzvah, I helped teach at the weekly Hebrew school and later became involved in a Zionist youth movement, both of which added to my consciousness of Hebrew. Formal language learning only began for me when I went to secondary school. Educated in a highly selective and elitist system, I was the only boy from my local primary school to go to an academic high school. There, because of the vagaries of the new curriculum, I was required to learn French and given a choice between mathematics, German, or Latin. I chose Latin. In the sixth form, I was assigned to an highly selective scholarship class—of the 12 boys in the class, 5 became university professors; 2 went on to senior science positions; 1 became a

high school headmaster; 1 a doctor; 1 a banking executive; and 1 started out as a lawyer, became a judge, and ended his career as governor general—where I prepared for scholarship exams in what the form master considered a well-balanced program of English, French, Latin, and German.

NS: What are your early memories of being tested at school?

BS: My training in testing started in primary school with the intelligence tests which led to my selection for the advanced academic stream (I have memories of the satisfaction of filling in boxes), continued throughout secondary school with regular class tests, and then from the third year (fifth form), the end of each of the next 7 years was celebrated by taking an external examination after which I emerged with an M.A. and excellent exam essay skills.

NS: What subjects did you study at university?

BS: At the university, I continued my language interest, majoring in English, minoring in French, and failing first-year German. Our Scottish-born professor of English was deeply interested in language and proud of the fact that a good number of his students finished up as linguists or language scholars, including an editor of the *OED* [*Oxford English Dictionary*] and professors at Oxford and Cambridge. Unprepared by my education for anything else, I decided in my last year at university to become a secondary school teacher and spent a year in a teacher training college to become certified.

NS: So after graduating from teacher training college, what was your first job?

BS: My first job—I was hired, it turned out, not because of my teaching subjects but to coach the school field hockey team—was at a high school in a small town on the east coast of the North Island of New Zealand. Here, my teaching of lower forms (and the least academic of these) and my hockey coaching brought me in contact for the first time with many Maori pupils. Trying to teach them English (for which my university training seemed to offer no assistance), I became interested in the linguistic problems involved. I was struck by the fact that those who reported that they spoke Maori at home turned out to be better English writers than those who said they spoke English at home. This eventually determined my fascination with bilingualism and my lifelong concern for the central problem of educational linguistics, the choice of medium of instruction.

NS: I suppose that, as a teacher, you will also have had dealings with exams or tests in school? Can you recall any formative influences in assessment which arose from your teaching experience?

BS: It was while I was teaching high school that I discovered linguistics as a field and persuaded the local bookstore to find me copies of two books by Charles Fries which showed me that linguistics can be relevant to education (for later views, see the *Concise Encyclopaedia of Educational Linguistics* that I edited or the *Handbook of Educational Linguistics* that I am

Downloaded by [National Institute of Education] at 22:54 02 May 2014

Downloaded by [National Institute of Education] at 22:54 02 May 2014

currently preparing). I had a second important epiphany while teaching there. An educational reform offered schools the possibility to exempt students in the lower sixth form from the university entrance examination. To do this, schools were required to rank all their sixth-form students and to determine at which point on the list they wanted students to be granted university entrance without examination. Students below the criterion point could take the examination, and it was assumed that the number who passed would show how believable the school's ranking was. For some reason (presumably, my junior status rather than the fact that I had studied Latin rather than mathematics), the headmaster asked me to prepare the ranked list on the basis of school marks awarded by teachers in the various subjects. I had already been aware from my own experience as a pupil that marks varied among subjects: in languages, the highest marks were usually in the 70s and 80s, while our science and mathematics classmates would regularly score 99% or 100%. Remembering this, I prepared two lists, one based on reported raw scores and the other based on scores standardised to a common average (I had to learn how to do this; I assume I had access to an adding machine). This first experience of examination statistics and of the problem of interpreting examination scores has remained a defining anxiety in my academic career.

NS: When did you first decide to travel from New Zealand?

BS: My activity during student years in the Zionist youth movement led me to decide that I wanted to move to Israel like a good number of my Jewish friends. Many of my non-Jewish fellow students were also planning to leave New Zealand—one finished up at Oxford, one at the University of British Columbia, another in the U.S., and three spent some years in Australia. My sister, by the way, stayed in New Zealand, where she developed a career as a journalist and is now a teacher of journalism. So, after 2 years of secondary school teaching in New Zealand, I set off on my way to Israel, spending a year teaching in a secondary school in Australia and another year teaching at a minor public school near London. This encouraged my interest in language variety—the boys at the school in England were striking in the way that they would switch from Standard English in the classroom to their various local dialects in the dormitory. When I got to Israel, I spent the first 5 months studying in a Hebrew ulpan (a school for intensive study of Hebrew), living with other immigrants with whom the Hebrew we were learning was our only common language. Failing to get a job in a secondary school—the schools I applied to wanted English teachers with greater Hebrew fluency than I yet had—I discovered that the Hebrew University was hiring teachers of English as a foreign language, and my career was launched. The following year, I was required to do army service. Noting my background, the army assigned me (after basic training) to a

staff position in education concerned with teaching foreign languages. Returning to the university, I decided that I needed more advanced training in linguistics if I was to continue to teach languages.

NS: So when did you move to North America?

BS: The move to North America followed this—although my primary motivation was in fact to pursue a young woman whom I had met at the end of a summer she had spent in Israel and who could not be persuaded to stay. She was a student at a university in New England, and the closest I could get easily was Canada, where I had no immigration problems. My original plan was to teach at a secondary school in Montréal while studying linguistics at McGill University. Again, I had to change direction: There were no jobs available in the Protestant school system, so I had to fall back on a position teaching English at McGill University; and McGill University had not yet discovered linguistics. So I worked on my doctorate at the University of Montréal. But the main part of my plan worked—my wife and I were married in my second year in Canada, and she transferred to McGill to finish her B.A.

NS: And what encouraged you to move to the United States?

BS: As I was finishing my dissertation, I applied for a postdoctoral fellowship at the 1964 Linguistic Institute to be held at Indiana University—a great experience as the teachers that summer included Chomsky, Halliday, Weinreich, Hockett, and Haas, and as sociolinguistics was being invented down the hall. I was awarded the fellowship and offered a position in the Linguistics Department, which I accepted. At the same time, my wife started on her Ph.D. in English literature—she has since added interests in cognitive studies and art and continues research, teaching, and publication in the resulting pioneering field of cognitive criticism.

NS: What about your children?

BS: Our two children, one born in Bloomington and the second in New Mexico, rejected our academic bent; each, after service in the Israeli army, has taken degrees and pursued careers in computers. Our son has a small successful software company in New York and a major reputation as a blogger and software guru (try putting the family name into Google and he dominates the returns) and our daughter stayed in Israel—she has five children under 13 and is vice president of a software company here.

NS: How did you first get involved with language testing in the U.S.?

BS: At Bloomington I had two administrative responsibilities in my new position: to direct the English for foreign students program and to direct an M.A. program in the teaching of English as a foreign language. It was in the first capacity that I started serious research and practice in language testing and the second that I developed my interest in bilingualism, sociolinguistics, and ultimately language policy.

Downloaded by [National Institute of Education] at 22:54 02 May 2014

NS: And so was it from this cross-disciplinary interest that you first became involved in discussing testing issues with other applied linguists prior to setting up LTRC [Language Testing Research Colloquium] in the 1970s?

BS: Yes. We are talking now about the prehistory of LTRC and the group of language testers that started meeting almost every year. The first I attended was the 1968 Michigan meeting run by Jack Upshur—and then the same group kept on meeting at other places. I remember a meeting at Idlewild at USC [University of Southern California] which Eugene Brière conducted a year or two after that, and we met at Georgetown one year. At some point we plugged into TESOL [Teachers of English to Speakers of Other Languages], and then the meeting turned into LTRC as we know it. But the early group was discussing all these same questions that we're still discussing, such as functional versus formative tests. I remember a wonderful discussion with Jack Upshur one time—"Wouldn't functional tests be the best thing? Give someone the money and send them out to get cigarettes, and if he comes back with a pack of cigarettes, then we'd know he knows enough English. Or see if he can get a date with a girl." Jack replied, "If he's got the money, they'll give him the cigarettes even if he doesn't speak the language, and if he's got a car, he'll get the girl whether he speaks any English or not!" These are issues which we are talking about today, essentially how you distinguish the context and the task and language ability.

NS: Your first paper on specific issues in language testing was the 1968 paper, was it?

BS: Yes, it was on the noise test.

AK: Was this the reduced redundancy test?

BS: Yes, it was one of these wonderful accidental developments. Bengt Sigurd, a Swedish phonetician, was on sabbatical leave at Indiana University, which had a wonderful weekly seminar on linguistics. They invited speakers in every field. We went to one meeting together—I don't recall what the lecture was about, perhaps information theory—and when we came out, we came up with this idea that it should be possible to measure somebody's knowledge of English by adding noise to a taped voice to the point at which they couldn't understand. That was the thought. It seemed to us that we would then have a very practical measure of somebody's knowledge of language. The trouble was that we didn't have the technology at the time to add noise to a tape in a fixed ratio. So we left it for a bit until I happened to go to Pensacola, where I visited a friend of mine who was in the Naval Air Station there, a doctor. He took me in to see some of his colleagues in the audiology clinic at the Air Station. They were studying the problem of increasing deafness of older pilots. The intriguing thing they noted was that older pilots didn't understand less when using the radio because they had had more experience of the possible messages than younger pilots. It

Downloaded by [National Institute of Education] at 22:54 02 May 2014

Downloaded by [National Institute of Education] at 22:54 02 May 2014

seemed that the increased knowledge of content balanced out the decreased hearing ability. In our case, in a sense we were arguing that the noise would balance the knowledge of the language. They had the technical equipment and prepared some tapes for us. So we started trying out the tapes, and they did a wonderful job of testing. What the noise test did most clearly was distinguish a good second-language speaker from a native speaker. As a practical measure, the test was obviously limited by the sentences that we had made up and recorded. Secondly, there were a significant number of test takers who were terribly upset and showed terrible testing anxiety in the noise condition, and so the test turned out to be impractical, though some people continued research. For me it was significant because it supported the theory of reduced redundancy.

NS: Perhaps it was a characteristic of those trait-based tests as Lyle [Bachman][1] has recently called them, such as *cloze* or *C-test*, which were rather threatening to the candidates?

BS: Yes, they could be. Of course, Jack Carroll never believed in the cloze as a language test because he thought the cloze was a separate ability. He wouldn't agree that it was a test of language ability. So it could be that handling noise was also a separate ability.

NS: And not obviously related to a task in a real-world context of language use?

BS: Yes, although one researcher did later on try to contextualize it by saying, "Imagine you're on an aeroplane, and this is the announcement"—that sort of thing—to try and give a context in which it would be more acceptable. I think it's still worth looking at in such things as tests of air–land communications. There's a lot of concern about aircraft crashes, often associated with miscommunication between the ground and the air, particularly now that everybody is using English and so many people are working in a second language.

NS: And not under ideal communication conditions …

BS: Right, and with increasing anxiety: "There's a plane in front of me, what do I do?"

NS: So again it's sort of pointing you in the domain of specific purposes again. So that was in the late '60s, beginning of the '70s. But in the meantime you'd also become involved with ETS [Educational Testing Service], in some way?

BS: In the meantime, what happened was that, once I got to Indiana University, I was connected up with the field. Tom Sebeok, who was one of the big gurus there and a highly organised busy professional, was a fantastic linguist, and a magician. He saw magic as part of semiotics. His favourite example was Clever Hans. Remember Clever Hans—the Austrian horse who could

[1]On the Thursday morning before the interview took place, Lyle Bachman gave a plenary talk entitled "What Are We Assessing? The Dialectic of Constructs and Contexts in Language Assessment."

Downloaded by [National Institute of Education] at 22:54 02 May 2014

do arithmetic? It was a circus performance; you'd give the horse a mathematical problem, and it would tap its foot a number of times and give the correct answer. People tried to work out for years how it did this. It turned out it wasn't doing mathematics. It was just watching the faces of the people around and when it got to the right answer, it recognised the changes of expression and stopped tapping! That's real magic—a pretty highly developed skill that nobody recognized. Anyway, Tom got me connected with linguistic activities. He got me an invitation to the meeting where Harold Allen reported on the first U.S. national survey of the teaching of English to non-English speakers in the States. And out of this initiative developed TESOL and various other organisations. So having been at that meeting, I got to know people connected up with the growth of TESOL when it was founded, and having been at Upshur's Michigan testing meeting, I got to know language testers, too. As a result, I was invited to be a member of the Committee of Examiners for TOEFL [Test of English as a Foreign Language] at what was an interesting period. It had by then lost its independence, but it was still jointly owned by the College Board and ETS. The usual setup at ETS is for a test to have a nominally independent board, appointed by ETS staff, who would have responsibility for the test and contract the work to ETS, who could say that there was a board setting policy. But in this case, the College Board was still taking a serious interest in TOEFL and so everything about the test had to be justified to the College Board's people—people I knew through the National Association of Foreign Student Advisors [NAFSA]. I was active in the Association for Teachers of English as a Second Language, who were the English teachers in NAFSA. That gave me a political connection with people who were involved in running TOEFL at that stage. I was appointed to the Committee of Examiners and spent 2 or 3 years being invited to Princeton and put up at the Princeton Inn (which an assistant professor could not dream of affording in those days), driven out to the campus, and met all the serious senior people at ETS like Bill Angoff, with whom I had many fine conversations. I remember one trip we made one day (it was snowing) as we drove from the College Board's office in New York out to the ETS campus at Princeton and the two of us were sitting in the back of the car discussing language tests. The particular topic was the very high correlation between parts of a language test—that whatever you did, you were getting much higher correlations than anybody did in any other kind of test. The exception was Latin, the one language that acted like tests in other subjects in which the parts were reasonably independent. Of course, this supported our thoughts about overall language proficiency and that sort of thing. So talking to the researchers at ETS was very exciting. But then we sat down with the test development people who would take us through the printouts of results.

Downloaded by [National Institute of Education] at 22:54 02 May 2014

NS: Who were they?

BS: I don't know. They were just technicians; they weren't testers; they weren't language testers—nice people, but they just got the printouts. There was nobody in language-testing research working with TOEFL—the woman who was running it was an assistant manager who had been put in charge of TOEFL after Les Palmer left. We'd get to meet researchers like Angoff and John Clark, but they didn't sit with the committee. I remember spending one wonderful day going through the whole of one part of the test being unable decide what the items were testing—were they vocabulary or idiom or lexicon or grammar or what? I think, at the end of the day, we realized that the separate vocabulary test, with all the problems of memorization, was not needed.

NS: Were you involved at that time in the discussion about whether to have the writing test?

BS: No, that wasn't an issue. That didn't come to us. We were the Committee of Examiners. There was another committee somewhere which discussed policy matters. The two committees met at different times. What I think embarrassed ETS is that some of us on this Committee of Examiners knew more about testing than we were supposed to—usually, the Committee of Examiners consists of subject matter experts who have no idea of what's going on in testing.

AK: Yes, I agree with you because I was on the Committee of Examiners Board for 6 years, and we were a group who knew a lot about language and a fair amount about psychometrics, too, unlike the mathematics professors who were brought in to design the mathematics section for the GRE or the SAT. They knew the mathematics part, and they seemed to rely more heavily on ETS experts to give them expertise in terms of testing. But in our case we were fairly knowledgeable.

BS: Well, the mathematicians could have argued with the statistics, but they couldn't argue with the psychometrics, but we could in lots of ways.

AK: In your book, for example, in *Measured Words*, you were critical of the use of expertise from the TOEFL committees. You said something like ETS would invite well-known language testers to be on the Committee of Examiners and use that as a way of claiming that they have consulted with the best.

BS: Well that's the basic ETS structure which I learnt about when writing that book. Remember the way in which ETS was set up: The College Board used to do its own testing and had an office in Princeton for the design and production of tests. It decided, in the late '40s I think, to split the office off and set up ETS as an independent corporation to handle test production. The College Board would contract with ETS to develop and administer its tests. ETS was set up as a nonprofit corporation in New Jersey incorporated

Downloaded by [National Institute of Education] at 22:54 02 May 2014

under the New York Board of Regents. Now, a nonprofit corporation didn't have to report its finances to anybody. You couldn't find out what happened; you couldn't find out where the money went or where it came from. Most of it seemed to have gone into building the campus and the Chauncey Centre, not paying terribly high salaries at that time but making life very comfortable for its senior people, building up a wonderful research group who were given comfortable facilities on the campus and so on. The Chauncey Centre was very, very useful because it was open to senior testers in the country who wanted to come and study, and it was also available to appropriate senior government people who wanted to use it. The Secretary of State, as a regular thing, would have meetings and sessions there, so ETS had excellent relationships with government. In effect, ETS was virtually out of control. The other testing corporations, like Psychological Corporation, set up early in the '20s, were set up as for-profit organisations, and they were later taken over by publishers. But ETS remained completely independent and was completely untouched until the court cases around about '79, '80—when people actually got court orders to open up the exams and to find out what was going on. I remember a meeting of the Center for Applied Linguistics [CAL] Board of Trustees: William Turnbull—who had been vice president of ETS when it took over TOEFL and whom I succeeded as chair of the CAL board—was at the meeting the day that the court decision came out, and he was getting telephone calls from ETS all the time. Everybody was terribly nervous: "What's it going to do to our whole system if we have to actually let people see our exams after they've been given?"

AK: It was called "truth in testing" legislation.

BS: Until then ETS had been virtually untouched by anybody. There's a fantastic report done for Ralph Nader and his consumers organization which was investigating ETS operations. It was completed in 1980, and you never read such a frustrated piece of writing, as the investigator realises he can't get answers to his questions. He can't find out how they were being run and what they were doing, where the money was going and how they were making their decisions … a very interesting way in which institutions take over.

AK: In *Measured Words*, you said that both at ETS and at Cambridge, something like, people made decisions that were not professional or academic but somewhat personal … so that they would champion their own ideas.

BS: Well, institutional, political—but if you look at any organisation, it is like this. There were different struggles going on in the two organisations. The thing that I looked at particularly in the book was the way in which TOEFL was taken over by ETS, which was a very intriguing event. Originally, the

initiative for a test of English as a foreign language came from the U.S. government—after the passing of the 1924 Immigration Act, which, you remember, set quotas for immigrant by picking years in which the right people came.[2] They picked the years when there were few Orientals, Slavs, Jews, or Italians but large numbers of Northern Europeans. That's how the quotas were set up. This reflected the growing isolationism that developed in the U.S. after the First World War. Psychologists have some responsibility for this. Brigham, who had been involved in the mass army IQ tests, wrote a book which explained why Italians, Jews, let alone Blacks and Orientals, were not intelligent. He later changed his mind before he got his job at the College Board. Before that, he gave evidence to Congress on the Immigration Act. Shortly after the act was passed, the immigration authorities noticed that there was a loophole, as the act said anybody who applies for a visa to study at an American school (which meant high school or university) was automatically granted a student visa. How did you know they were actually going to go there and study? In 1926 it was realized that an English test would exclude applicants unprepared for study. The College Board developed a test, first administered in 1930 to under 30 candidates (the changed economic situation meant the numbers interested were low). It was given again in 1932 to 139 candidates, including 82 engineering students in Moscow, with testers sent from the U.S. The same test was given for the fourth time, in 1935, but no money was available to develop a new form. In 1938, someone suggested using it for groups of Jewish refugees who had to prove knowledge of English to be admitted to the U.S., but it was no longer available. So they had to rely on what else they had got,

[2]Immigration Act of May 26, 1924, 43 Stat. 153. In response to growing public opinion against the flow of immigrants from southern and eastern Europe in the years following World War I, the U.S. Congress passed the Quota Act of 1921 and then the even more restrictive Immigration Act of 1924 (the Johnson–Reed Act). In conjunction with the Immigration Act of 1917, this governed American immigration policy until 1952 (see also the Immigration and Nationality Act of 1952).

1921 The Immigration Act limited annual immigration to 350,000, and quotas for each nationality were introduced.

1924 The National Origins Act imposed a total quota on immigration of 165,000—less than 20% of the pre–World War I average. It based the number of immigrants allowed in from any particular nation on the percentage of each nationality recorded in the 1890 census. It was a blatant policy to limit migration from southern and eastern Europe, which mainly occurred after that date. For example, in the first decade of the 20th century, an average of 200,000 Italians had entered the United States each year, but with the 1924 act, the annual quota for Italians was set at less than 4,000.

1927 The annual immigration ceiling was reduced to 150,000.

1929 A revision to the National Origins Act was introduced. An immigration ceiling of 150,000 was made permanent, with 70% of admissions reserved for those coming from northern and western Europe and 30% reserved for those coming from southern and eastern Europe

Downloaded by [National Institute of Education] at 22:54 02 May 2014

Downloaded by [National Institute of Education] at 22:54 02 May 2014

whether they used the Cambridge test or what they used I don't know. There was a second effort to develop a test for foreigners after the Second World War. Charles Fries from Michigan was invited to a planning meeting and took his student Robert Lado with him. The College Board then developed a test of English as a Foreign Language that they used for quite a few years. That test was used but was not secure, and in the late 1950s, the demand for a more secure test arose. Charles Ferguson, with Ford Foundation support, called a meeting in 1961 at the Center for Applied Linguistics, inviting John B. Carroll and Robert Lado and other testers. That is where TOEFL was designed, planned to be run by an independent organisation. Two experienced language testers, David Harris and Lesley Palmer, were given a year to develop it. They were assisted by testing experts from ETS with funds from Ford. As the year went by, it was clear that it was going to take longer than they thought to do it. They had set up a test committee made up of experienced EFL [English as a foreign language] testers and teachers from all around the country to become test writers, working in pairs at the same school, each pair given a section of the test to write. The committee was brought in for a week to discuss the specifications for the items and then sent home to write. When the items came in, David Harris and Les Palmer started editing them. Now I don't think a single item sent in was acceptable, so Palmer and Harris spent all their time writing items. As time went on, the money started to run out, so they went back to Mel Fox, the person at the Ford Foundation responsible for all the language-related grants. In preparing a case to present to the head of Ford, he visited the College Board and asked the vice president to seek the board's support for this. The detailed story is in my book *Measured Words*. In the College Board records, I found his papers, and among his papers was the draft of his statement to the board, in which he said, "I don't know why we should support this project. I don't know these people. I don't know anyone working in it." This was a lie: The vice president had worked closely with the ETS expert on a number of tests. So the College Board simply refused support for continuing the work. That upset Fox—he didn't know about the VP's treachery—but even worse was to follow. When Fox went to see the president of Ford, he started asking him questions like "Why do you need a test like that in English. Why don't you give these people tests in their own language?" Fox was surprised at the question. But it was the time when the College Board was planning the Spanish SAT, so it was pretty clear someone from the board had talked to the president of Ford. There's a whole lot of earlier correspondence that explains the College Board's nervousness. They sent a junior person to the meeting at CAL in 1961. Directly after the meeting, she called the New York office and wrote an urgent letter to the president of

College Board, who was on sabbatical leave in France, saying that a terrible thing was happening: A new group was trying to take over EFL testing and establish their own exam. In his presentation later to the board, the treacherous vice president was even more specific—he saw it as the newly come Ford Foundation trying to muscle in to areas better left to the Carnegie-supported establishment. As a result, the president of Ford said to Fox, "No, they can't have that much money; there isn't the support for it." He went back to the TOEFL Committee and said, "We're sorry. We can't get any money." So the TOEFL Committee was in a quandary. Just by chance, the representatives of College Board and ETS who were members of the committee said, "Oh, we'll take it over for you. We'll handle it." So TOEFL lost its independence and was handed over to joint ownership of College Board and ETS. They, of course, went straight back to the Ford Foundation, who came up with all the money needed to do the development. But the story doesn't end there. TOEFL now belonged to College Board and ETS jointly—that's the period that I was on the Committee of Examiners. Now, during that period, the test continually, every year, lost more and more money, until finally somebody from ETS said to the College Board, "Well, you know, you're losing money on this. Do you want to get out?" and they said, "OK," and they got out. And the next year, with ETS as sole owner, the test stopped losing money, and it remained under ETS control from then on and was one of the most profitable tests they had for many years.

This is where the comparison with what happened in UCLES [University of Cambridge Local Examinations Syndicate] is intriguing. Once ETS had the test, they simply kept it going as part of their machinery, and when Les Palmer, who was a director the first year or two, left, they appointed a woman who was an administrator, a business manager from somewhere, to run it. And that's how they continued to have it run, by a business manager. They responded when there was external pressure, like the pressure for the speaking test which was developed when the state legislators started complaining that their children couldn't understand the foreign assistants who were teaching them math and computers. The writing test was also a response to external pressure. But it was always from outside; it was never internal.

The intriguing difference with UCLES was that after it had gone through its big shake-up as a result of the comparability study and the realisation that the test was not psychometrically defensible, it set up a new organisation, an organisation that was able to keep putting the money back into test development, presumably partly because it was an embarrassment to UCLES to make money. At the time of the TOEFL–Cambridge comparability study

Downloaded by [National Institute of Education] at 22:54 02 May 2014

(Bachman, Davidson, Ryan, & Choi, 1995),[3] they had just given £2 million sterling to set up Gillian Brown's new centre.[4] When I asked why, I was told, "Well, we've got this money, and if we don't spend it, the university will take it from us anyway." UCLES couldn't hide its money.

NS: Yes, as part of the university the only "shareholder" of the syndicate is the university itself. It can use surpluses for educational good works, including, of course, reinvesting in itself or other aspects of the university. The endowment for Gillian Brown's department (which is now John Hawkins's department) was one of those. And so it had a different history from ETS—and up until the 1980s, it didn't actually have any surplus money.

BS: Yes. If I remember rightly, the whole of the comparability study was to try and keep control of the European market when Cito was threatening to start its own version of TOEFL which could have become a competitor. So it's nice to have an economic issue turning up. But then comes the critical question, what *do* you do with profits? ETS didn't attempt to build up its test, but UCLES did, and from that point of view, I suppose UCLES wins in a very real sense.

NS: It's interesting though that your recollections take us back to 1961 and, then, to when TOEFL became established—when was it, 1964?

BS: Well, at the end of the first test, yes.

NS: And by that time, had ETS decided not to have test of writing?

BS: They decided that at the 1961 meeting. Again, there was this paradoxical mix of motives. The ETS man who helped plan TOEFL as a testing expert, at this stage in 1961, was working with the vice president of College Board (the one who denied knowing him). They were trying to get writing back into SAT and were conducting experiments to show how to do it. But when he came to the TOEFL meeting in which they were planning the new test, in like 2 minutes of discussion, somebody says, "Well, can we have a test of writing?" He replied, "No, it's too expensive. You can't afford the cost of airmail with Pan Am. We'll never afford postage." End of discussion of a writing test.

[3]This was a major study which compared the Cambridge First Certificate in English with the Test of English as a Foreign Language and investigated similarities in test content, candidature, and use. It was carried out between 1987 and 1989 by Lyle F. Bachman, Fred Davidson, Katherine Ryan, and Inn-Chull Choi on behalf of the University of Cambridge Local Examinations Syndicate. It is described in full in the first volume of the Cambridge English for Speakers of Other Languages / Cambridge Univeristy Press Studies in Language Testing series (1995) and an extension study in Kunnan (1995).

[4]The University of Cambridge's Research Centre for English and Applied Linguistics was established in 1988, funded by an endowment from the University of Cambridge Local Examinations Syndicate. It is a freestanding department of the university, though its staff members are also members of the faculty of English. The first director until 2004 was Professor Gillian Brown. Since then, the post has been held by Professor John Hawkins. See www.rceal.cam.ac.uk

Downloaded by [National Institute of Education] at 22:54 02 May 2014

Downloaded by [National Institute of Education] at 22:54 02 May 2014

There was another administrative bureaucratic matter. One of the reasons the test, in fact, was not making money was the understanding that, after candidates paid the fee to take the test, they could have as many reports as they liked sent to the universities. The cost of preparing and mailing a report was high, and that's where all the money was going—that's why there was no money to include writing—certainly no money to do any work on a speaking test and no money to think about the predictive testing, which they originally thought they were going to do. No money for research.

NS: So there wasn't a principled rejection of the more subjective in favour of the objective testing which had come out of the structural movement—the Lado and Fries era?

BS: No, it was because the subjective was expensive.

NS: And not because of the principle or because it was unreliable?

BS: Yes, that's right. There was plenty of relevant research going on at ETS. There were plenty of people developing new kinds of testing techniques. But it was never under the control of, or even directed towards, TOEFL.

NS: From what you describe, TOEFL, when up and running, developed its own momentum for 40 years.

BS: Yes, you couldn't change it. I mean, you couldn't attack the standards because the test had been calibrated on that very first group. And the first group was a fairly unnatural group of students already in the United States, studying at larger universities, and that's where the calibration was done and from then on everything had to continue to agree to that. So with that on the one hand and with the enormous growth of candidates from Hong Kong, Taiwan in particular, and then Japan and with the ways that they prepared for the test, you produced a test that was getting more and more meaningless to more and more people but which was absolutely established. Then TOEIC [Test of English for International Communication] was developed as a new way of selling TOEFL. It was good business.

NS: When did you first think about your three-trends analysis, then, of the eras, if that's the right word, of testing?

BS: Well, I suppose that was fairly early on, a self-conscious concern about "Where were we?" and "What were we doing?"—and it was quite wrong, now that I think about it. It was one of the reasons that in the measurement book I got into history. I suddenly realised that I'd written this article with the three trends, which was the way we saw things when I started out. There had been something traditional in the old days, the psychometrist–structuralists came along and fixed everything, but now we're sociolinguists and transformationalists, and all that stuff must be wrong, and we're doing the really good stuff, right? We're the "modernists," or whatever it was. But then, when I started writing about testing history and looking at it more closely, I started wondering, "What really was going on?" I

Downloaded by [National Institute of Education] at 22:54 02 May 2014

wasn't at CAL in '61, you know—actually, I was in the Israeli army at the time. I wasn't in America, and I didn't know what was happening. And I came in with a view of the field at that time that assumed that everything after '61 was good and that everything before then was bad. I had terrible fights with Lado, not with him personally but with his students and his colleagues who got terribly upset at the things that not just me but others would say about him, and really we were quite wrong. We misread Carroll. We assumed Carroll had said that Lado was wrong, but what Carroll said was, seeing that Lado has just written this wonderful book all about item tests, he would now talk about what's left over—integrative testing. And, also, the other problem is that Carroll *had* written a history of language testing that I finally found buried in the archives. It was in '54 I think. He was gearing up for various testing studies, and he gave a paper at the Georgetown Roundtable and offered the manuscript of his state-of-the-art study to Georgetown to publish, but they said, "No, we don't need it; we don't want it." There are half-a-dozen copies around in a few libraries. It is really very good and a clear statement of what people knew about language testing at that stage. Well, it worried me—that I really hadn't learnt all that real history, so I tried to go back and read what I could. And *Measured Words* got longer and longer. The first half of the book is leading up to 1961, aiming to say this is where it was all coming from, this is what was going on. There was a lot of very exciting research and development going on, you know, with somebody like Jack Carroll, really brilliant, wonderful. So I got a much different view as time went on, and my book tried to develop it. I'm still rethinking, as one does, about what goes on. There's a sort of critical break that hasn't hit us as hard as it should, between, what shall I say, the people who base everything on linguistics and the linguistics that is going to relate ultimately to the structure of the brain, and so on, the sort of physical embodiment of language ability and the breakaway by the sociolinguists, the people who want to fit everything into a social context. Again this is presumably what Lyle [Bachman] is trying to clarify, but he hasn't gone far enough in either direction. But it's fascinating because both of them are there; in other words, language exists in the brain, and the brain is ultimately chemical actions, but the shape it takes depends on social structures of a very complicated kind, and you build up a very complicated construction of all the things working in together.

NS: And it's never fixed?

BS: And it's never fixed. And it never stops, and it's always moving, and they are all variable, and your chemicals may be getting mixed up all the time! And anyway, most people don't understand other people most of the time, even speaking the same language. Not quite perfect communication either, so why should you be able to test somebody? But anyway, what was going

Downloaded by [National Institute of Education] at 22:54 02 May 2014

on here with me, I now realize, is that while my first interest in language testing was, I suppose, much more linguistic—"What does it mean to know a language?" I asked. But hearing Bob Cooper's paper at the Michigan meeting in 1968 where he brought the sociolinguistic aspect in for the first time—he'd read Dell Hymes, *The Ethnography of Speaking*, which wasn't really published widely yet—and knew about the start of sociolinguistics. His Michigan paper was about the testing Fishman and Cooper were developing for the barrio study in New Jersey, and they realised that it wasn't just enough to measure the four skills. You had to deal with those skills in their various social contexts, and this is where Fishman's notion of domains moves in and you start getting a fascinating sort of ecological picture of the person, profiles of quite complicated differences in people using different languages in different domains, being better in different domains with different things. There are, for instance, plenty of people who do all their speaking in one language and all their writing in another. So I got more and more interested as time went on in this dimension of things and moved away from testing and moved into language policy. Well, first, educational linguistics and into language policy, now into language management. I remain connected with language testing, but it is no longer my central concern.

I remember the study we did when I first started to work with the Navajo Nation. I moved to New Mexico in 1968, and we started the Navajo reading study a year later.

AK: And you published a paper on Navajo language maintenance in 1970 …

BS: Yes, so that was the beginning of the project. Someone came along to us from the Bureau of Indian Affairs and said, "Here's a research project for you." I'd just arrived in New Mexico, from Indiana, where I had become interested in bilingualism but nobody was bilingual except foreign students. In New Mexico in 1968, the students were marching up and down outside the university, saying, "We want Spanish back. Give us back our Spanish." But there were plenty of my Spanish-speaking colleagues to work on this. A bureaucrat from the Bureau of Indian Affairs explained their problem. He explained that 50,000 of the 120,000 population of the reservation were in school and suggested I look at them. I went out and saw a huge boarding school with a couple of thousand students. I went from one class to another, and in every classroom I was taken into, the teachers were standing up talking in English and the kids were sitting there not understanding them because the kids didn't know a word of English. And so the bureau official asked me if I'd be willing to replicate the Modiano study from Chiapas in Mexico which showed that if you teach people to read in their own language first, they learn to read the standard language faster. So I said, "On condition that we can find out, first of all, if the language is still there."

Downloaded by [National Institute of Education] at 22:54 02 May 2014

They agreed, and so we made a survey of 6-year-old children coming to school for the first time. We asked the teachers to say, when the children started school at the beginning of the year, "Did they know any English or not?" Of course we wanted to validate the questionnaire. We were a bit nervous about using teachers who didn't know Navajo. So we needed to design a test to see if 6-year-old Navajo children knew their language. We came up with a very simple and intriguing test. You simply say to the child in Navajo, "Who are you born for?" This is a traditional greeting: "What's the name of your parents' clans?" Traditional Navajos ask this question to avoid incest problems. Anybody who answered that question came from a traditional home where they still spoke Navajo. So you were in no doubt who knew Navajo. If they couldn't answer it, then we tried to talk to them in Navajo, asking a set of simple questions. For the English validation, we had another set of simple questions. It was the combination of cultural and linguistic knowledge, which was intriguing. Later, the study moved on the sociolinguistics, the language shift, the whole issue of literacy, all questions that moved me quite a bit away from testing generally.

NS: But the "use with care" idea with testing, presumably that came out of the interest in some of these policy and language use areas?

BS: Yes. That came out of an invitation to IUS in Germany. Remember IUS? It was one of these groups in the early days —the Interuniversitatsprachtestgruppe (Inter-university language testing group).

NS: Was that in Berlin?

BS: No, it was Duisburg. A small group of testers there—Doug Stevenson, who had studied with me in New Mexico; Christine Klein-Braley, involved in developing the C-test; and Rüdiger [Grotjahn]. They started a little LTRC as it were, a little testing group, and had several meetings. The first time I was invited, I'd been to Germany once before, but going to Germany was a very difficult thing for me. I'd been to Stuttgart—I gave the test history paper, but I had left after giving my plenary—because I just couldn't stay there. So I was on the way to Germany again and trying to think about what to say at this meeting. In New York, on the way, we went to Ellis Island and were told about Ellis Island, and in particular I was struck by the staircase.[5] When the immigrants came off the boats, they came into a huge hall with a large staircase leading to the examination rooms on the next floor. There were inspectors on the staircase with chalk: If they saw anybody having trouble going up the stairs, probably indicating lung or heart trouble, they put a cross on their back, as a signal to send them back to Europe. They

[5]In 1890 the U.S. government selected Ellis Island as the Federal Immigration Center for New York. The centre opened in 1892, and when the immigrants arrived there by sea, they were led through inspections to gain entry. The inspectors examined them for any signs of illness and asked questions as to their origins, where they were going, and their suitability for work in America. About 2% were turned back.

Downloaded by [National Institute of Education] at 22:54 02 May 2014

were administering these rapid "medical tests" to make sure that only healthy immigrants were admitted. If you had enough money and you weren't travelling steerage, you didn't go to Ellis Island. You had your interview on the ship and were landed at the port, but if you came steerage, you went through this staircase examination. And so putting that together with concerns about tests being used to make, I don't know that we called them *high stakes* in those days, whatever it was, tests being used to make decisions like that, I gave a paper on the ethics of language testing.

AK: This was the paper titled "Some Ethical Questions About Language Testing" in the Klein-Braley and Stevenson collection.

NS: You had already reached some conclusions about the limitations of tests apart from their uses, the nature of testing scales and the reliability of tests, the notion of "unavoidable uncertainty." It had already occurred to you as a major problem in language tests or tests in general?

BS: Of all tests … Edgeworth's argument is that it's not clear that human abilities are measurable. They are *judgeable,* but judges differ, as they should. I keep quiet these days because I'm still very uncomfortable about training of judges.

AK: In fact, yesterday after the presentation on the raters, I was uncomfortable, too, about this movement to make raters totally uniform … I felt what was being proposed was that raters have to come to the rating table, as it were, after removing their backgrounds, their own understanding of the world; their own preferences and interests have to be left behind, and only the views that are usable for the rating which is set up by the test developer or the scoring rubric people are to be available. And my thought was that, if we did this, then we'd have people on the rating panel who are all identical judges. Do you have any thoughts on this?

BS: Well, you have a technique—having two judges gets you reliability it appears.

NS: And some triangulation of views …

BS: Yes, and that's enough. But if you train judges so that they agree, then you finish up with one judge. You've got good arguments that if the judges are all the same, then you only need one judge.

NS: But in practice you don't get them to agree perfectly—they only agree within limits because humans have a tendency to bring different perspectives to bear. So should we reach the conclusion that .80 reliability is probably what you want?

BS: As much as you want in judgement. But having said this, then you know you are putting enough doubt into the final result that you have to be very, very careful with the use of the final results. That became the critical issue and the fact that the kinds of decisions which have been made usually, the decision point is at the very point at which the maximum variation is going on, right? If you simply cut off the top or cut off the bottom of the group,

Downloaded by [National Institute of Education] at 22:54 02 May 2014

that is OK, but if you try and put a cut anywhere in the middle, then you know you are dealing with areas where there really isn't too much difference. So it becomes a problem. The other big issue, I suppose, again looking at this now more from a language policy–language management point of view, is the amount of public and bureaucratic ignorance about the meaning of the test scores that, presumably, somehow, we must also do something about. I had some very hairy sessions with Israeli ministers of education when I was chairman of the ministry's English Advisory Committee on the inadequacy of the system by which their tests operate—no pretesting, no item analysis, some sort of quick monitoring and then broad interpretations. "The scores are better this year," "They are worse this year," and this kind of stuff. I remember sitting with this minister of education and trying to explain to him what calibration means, and he said "No, don't be soft. I was a teacher. I know what happens. If you want to give a test, you write down some questions. You give them to the boys, and you see how many give correct answers, and you know all you need to know about them." Well, it's that kind of view of what is going on in the business—the occasional headline in newspapers: "Big concerns about level of education—50% of students on this test failed and scored below the average…" 50% below the average! It's the assumption that whatever testing system is there, is correct—the belief in the examiner's semidivine status.

NS: And perhaps a view that "if it was good enough for me, it is good enough for my children"—which has the effect of perpetuating the system.

AK: There's a notion in laypeople's minds that tests are infallible. So, if you get a low score, it must be because you didn't study hard enough. This is the case in many countries, and so test developers can continue to do what they are doing and not change or make things any better.

BS: One thinks about the SAT. Brigham kept saying, "This is not a test of intelligence, but it does seem to predict reasonably well how people will do at university. And that is all we kept telling you with this." And everyone said, "Ah, SAT depends on intelligence." And, of course, with the new testing movement—when was the last big one—in the '20s and '30s?—when tests really blossomed in America and didn't in England. In England, tests flourished in the 19th century—Gilbert and Sullivan in *Iolanthe* have this lovely couplet about selecting dukes "by competitive examination"—and then picked up again with the 11+.[6]

[6]The 11+ is an examination which was created as part of the 1944 Education Act and was given to students in their last year of primary education in the United Kingdom. The name derives from the age group of the students. It examines the student's ability to solve problems using verbal and non-verbal reasoning and was used to decide which type of secondary education would be suitable for students in a system with three strands: academic, technical and functional. It was largely discontinued in the 1970s but is still used in a number of counties in England and N. Ireland.

Downloaded by [National Institute of Education] at 22:54 02 May 2014

NS: Yes, the meritocracy movement under the Victorian era led to examination boards being set up—UCLES in 1858 and the Oxford Delegacy a year before that (see appendix).

BS: By popular demand … UCLES was not set up by Cambridge because Cambridge wanted it; it was set up by Cambridge because—who got the first one? I think a school in Exeter wanted a test—and they wanted a local test.

NS: It was a group in Exeter which led to the setting up of Oxford Delegacy. They wanted exams to be set by the university, but they wanted it to be taken locally. The same was true when UCLES was set up following petitions from other schools—hence, this apparently strange use of *Local* in the name Local Examinations Syndicate.

BS: It was by popular demand because people wanted tests.

NS: It's interesting. You now see a popular demand for tests in other European countries—perhaps it is a way of introducing a more merit-based approach to hedge the problems of nepotism and favouritism.

BS: Right, that's the kind of thing the Indian Civil Service Examination was about. Germany resisted public testing for a long time, but it had testing for magistrates in the 19th century. But they didn't like tests.

NS: And even now that persists.

BS: With the PISA [Programme for International Student Assessment] project,[7] for the first time they suddenly discovered that their people were not doing as well as they thought and they were thoroughly shocked—and they had to develop tests to find out what is going on.

AK: In the case of the Cambridge EFL examinations, the CPE [Certificate of Proficiency in English] started in 1913. Was that the exam which was used for the Indian Civil Service?

BS: No, there was no connection. The Indian Civil Service exam was fantastic. In the year of the Reform Acts (1833 and again 1853), Thomas Macaulay pointed out the value of tests: "Now look at X. He was a first ranker at Cambridge—he's now the lord chancellor. All these people who did well in these exams are now the top of the civil service, and we should do the same thing in India. If our language were Cherokee, any man whose language was Cherokee and could write poetry in Cherokee and who could pass Cherokee would

[7]PISA is an internationally standardised assessment that was jointly developed by participating countries and administered to 15-year-olds in schools. It was implemented in 43 countries in the first assessment, in 2000; in 41 countries in the second assessment, in 2003; and in at least 58 countries in the third assessment, in 2006. Tests are typically administered to between 4,500 and 10,000 students in each country.

"PISA assesses how far students near the end of compulsory education have acquired some of the knowledge and skills that are essential for full participation in society. In all cycles, the domains of reading, mathematical and scientific literacy are covered not merely in terms of mastery of the school curriculum, but in terms of important knowledge and skills needed in adult life." See www.pisa.oecd.org.

Downloaded by [National Institute of Education] at 22:54 02 May 2014

be the best person for all these jobs," etc., etc., in Latin and Greek, which it was. So they said, "Let's have an entrance exam for the civil service rather than let the governors continue to nominate their friends and family." The exam consisted of, naturally, Latin, Greek, history, geography, mathematics, Sanskrit, and a half-a-dozen other subjects, and everybody had to take everything. After the exams had been going for a couple of years, they looked at who had in fact been selected. Those who passed were either from Oxford, Cambridge, or Trinity College, Dublin. One Indian gentleman made it into the civil service—he must have got 100% in Sanskrit!

NS: The idea was that exams had to be hard, very difficult. This was represented in the first 1913 version of CPE, which I think lasted something like 13 hours of testing—and the subject matter would make you faint at the thought of answering any of those questions. CPE hadn't actually come from a requirement for selection, as TOEFL did. It had come from the changes in language teaching—the direct method and so on and the influence of the linguists and phoneticians like Daniel Jones. Daniel Jones became the first chief examiner of CPE and, of course, continued as one of the major influences in British linguistics right through to the 1940s. And so it was a different genesis, I think, than perhaps the TOEFL, which was set up with selection purposes in mind from the beginning. Although the idea was that, if you wanted to be a teacher in English, you had to reach a very high level and to submit to a very thorough and rigorous examination of your knowledge. So it was a high-stakes exam in a different sense.

BS: So now we are back to it again with standards and accountability and everybody believes that testing will solve all the problems which is why I tend to agree with Elana (Shohamy) that we probably don't spend enough time talking about the social, political and economic implications of what we are doing. That we are too much at the level of worrying about how to polish our instruments a little better rather than worrying about how they are *used*.

NS: And helping people to understand better perhaps how they should be used—or the limitations, shall we say, of their use. And this is going back to the kind of public service we owe to the stakeholders. In this regard, do you have any thoughts on codes of practice, such as the one ILTA [International Language Testing Association] is now developing?[8]

[8]ILTA developed a Code of Ethics in the 1990s, and this was followed by the drafting of the Code of Practice (which was discussed in Ottawa, 2005). The Association of Language Testers in Europe introduced a Code of Practice in 1994, and this has been developed into a Quality Management System (2006). Following the setting up of European Association for Language Testing and Assessment in 2004, the drafting of a Code of Practice began in June 2005.

Downloaded by [National Institute of Education] at 22:54 02 May 2014

BS: I believe that the ILTA Code of Ethics and Code of Practice are serious attempts to deal with this issue and can be key elements in the professionalization of language testing. It is probably too much to ask that we'll ever succeed in policing the misuse of language tests and the unjustifiable claims for their validity, but it is very important that we recognize and proclaim our responsibility. We are still not talking about these issues as much as we might be—talking about the way in which which tests might be used. This again goes back to my time at Indiana University, running this MAT (Master of Arts in Teaching) programme. The students were wonderful people who must have been pretty good in their own countries, who had come to the U.S. to study, and here was I, a beginning assistant professor, teaching them all about how to teach English and how to do testing and all the things they had to know, right? And towards the end of their 1-year programme, I would say to them, "What are you going to do when you get back and they would start talking about local conditions?" And I remember on a couple of occasions saying to somebody, "When you get back, don't become a teacher. You've got to go into politics because you've got to do something about that system to produce a place where you can do any of the kind of teaching we've been talking about."

NS: Exactly, I often find this when I'm talking to European colleagues who are finding difficulty in their own context in changing their exams for the better. These are people who know what should be done, for example, from a technical point of view. But actually what they need to do is to change the management of their organisation, to have some resources and have some support, and then they can do what they need to do. But just knowing what needs to be done from a technical point of view doesn't help you to change your tests or the quality of what you do.

BS: We had a recent example of this in Israel. We've had a major change in the Bagrut examinations in English (in Israel) over the last couple of years as part of modifying the exams to put in a new syllabus we've developed over several years. There is a writing passage in the test—people were told to write an essay and not to worry if they needed more paper but simply to ask for it. But none of the supervisors had any extra paper—it wasn't provided. All the candidates who asked for extra paper were told they couldn't have it. Some tried to write small or around the edges or things. The panic created within this system by this simple little bureaucratic decision—no paper!

NS: So, I suppose you would agree that one of the most significant trends in recent years has been the influence of sociological considerations on various aspects of language testing?

BS: Yes—as I said, it was Bob Cooper back in 1968 who pointed out the need to take account of social context. And with the development of communica-

Downloaded by [National Institute of Education] at 22:54 02 May 2014

tive language teaching and the influence of Hymes from the early 1970s, language testing has broadened from the academic testing of the standard written version of a language to allow for assessment of other varieties in various social and functional situations. One inevitable conclusion has been the realization that tests need to find some way to achieve authenticity, to measure the ability to perform in situations not unlike the real world. These are essentially questions about test use rather than form. There are important questions that need empirical resolution, such as "How does knowledge of one variety predict ability to use another?" and "Should we overcome our normal academic prejudice and try to teach and test dialects, pidgins and code-switching varieties?"

NS: In light of the greater awareness of the social context of language assessment itself, which we have already touched on, what do you see as the way forward?

BS: Recently, there has been an important adjustment of language testing to its social context and the attempt to define or judge the social context of test use and the ethicality of the test. We have already discussed my early papers about ethicality, but these thoughts have now become more focused, for example, in the work of Alan Davies, Liz Hamp-Lyons, Elana Shohamy, and Tim McNamara.

NS: So what does this greater focus on the social function of assessment mean for language testers in the 21st century?

BS: Well, we still need to retain a healthy scepticism about the validity of a test while recognizing the potential power of test results and the general need to use tests with great care. We also know that for practical, economic, political, and bureaucratic considerations, there are plenty of tests out there that continue to use unreliable and unvalidated testing techniques. Above all, there is a wide untrained public all over the world that believes language testing is easy and its results trustworthy.

Communication of the strengths and weakness of tests is our major challenge. As I have often pointed out, Edgeworth (in 1888) found evidence of the unreliability of traditional examinations and concluded that they were a kind of lottery in which the better candidate had the better chances. But they provided a way of sorting people out in ways which were preferable to the ancient method of casting lots for "honours and offices." He later concluded that whatever cannot be correct should be borne patiently; we should maintain a "liberal curiosity" to the sources and likelihood of error. I think the future depends not so much on what the testing researchers discover but on how effective they are at communication.

AK: You mentioned earlier the U.S. immigration policy of the 1920s. This has now become a hot topic again in many parts of the world. What are your

current thoughts on the direction of language assessment practices with regard to immigration, citizenship, and asylum?

BS: Many of us start our writing or talking about ethical uses of language tests with the biblical shibboleth test. We, of course, need to distinguish from the underlying policy (choosing "desirable" immigrants) and the use of language tests for this purpose. Just as psychologists made clear the inefficiency and inaccuracy of lie detectors in identifying lies and liars, language testers need to continue to point out the difficulty of interpreting language profiling as evidence of place of birth.

NS: I have just noticed the time—I think the sessions are about to start again, so we ought to draw this to a close.

BS: Yes, it's been fun—we could go on for hours—but there are some talks I'd like to listen to this afternoon.

NS: OK, so let me just ask you one or two closing questions.
So, the limitations of our current understanding of the technical aspects—the nature of the underlying constructs and the inevitability of measurements error—make communication about assessment a difficult challenge, as you have just pointed out. How do you think we can meet this challenge?

BS: A scientific colleague of mine once remarked that a successful scientist is one who either asks a good question or shows how to answer a good question. No one does both. The problem we face is something like the problem faced by those who are starting to understand global warning, not to understand the phenomenon but how to persuade politicians and the public to do something about it.

NS: I would say in our conversation that we have talked more about the people and events than about the technical issues in language assessment. What can the *Language Assessment Quarterly* readers take away from this?

BS: That in a century or so of research, we have a done a pretty good job of developing testing techniques and theoretical explanations but that we are little further along in persuading stakeholders how to manage the inevitable uncertainty of test results.

NS: Finally, after a long and distinguished career, which is being recognised here at this conference, you are still actively involved in writing and talking about language testing. What do you see as the main contribution you can still make?

BS: As I said, my work now is more and more concentrated in language policy. Language testing is a necessary component of language management, so I will keep up my interest in the field. I'll continue to rant and rave quietly about the dangers of tests as well as their usefulness and continue to remind the field about its history as well as its responsibility.

AK: Thank you Bernard—on that note, let's leave it there for today.

Downloaded by [National Institute of Education] at 22:54 02 May 2014

SELECTED PUBLICATIONS

Spolsky, B. (1977). Language testing: Art or science. In G. Nickel (Ed.), *Proceedings of the Fourth International Congress of Applied Linguistics* (Vol. 3, pp. 7–28). Stuttgart, Germany: Hochschulverlag.

Spolsky, B. (1981). Some ethical questions about language testing. In C. Klein-Braley & D. K. Stevenson (Eds.), *Practice and problems in language testing* (pp. 5–30). Frankfurt, Germany: Peter D. Lang.

Spolsky, B. (1984). The uses of language tests: An ethical envoi. In C. Rivera (Ed.), *Placement procedures in bilingual education: Education and policy issues* (pp. 3–7). Clevedon, Avon, England: Multilingual Matters.

Spolsky, B. (1985). The limits of authenticity in language testing. *Language Testing, 2,* 31–40.

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

Spolsky, B. (1997). The ethics of gatekeeping tests: What have we learnt in a hundred years. *Language Testing, 14,* 242–247.

Spolsky, B. (Ed.). (1999a). *Concise encyclopedia of educational linguistics*. Oxford: Elsevier.

Spolsky, B. (1999b). Standards, scales, and guidelines. In B. Spolsky (Ed.), *Concise encyclopedia of educational linguistics* (pp. 390–393). Amsterdam: Elsevier.

Spolsky, B., Sigurd, B., Sato, M., Walker, E., & Aterburn, C. (1968). Preliminary studies in the development of techniques for testing overall second language proficiency. *Language Learning, 3,* 79–101.

OTHER REFERENCES

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we test counts. *Language Testing, 17,* 1–42.

Bachman, L. F. (2005a). Building and supporting a case for test use. *Language Assessment Quarterly, 2,* 1–34.

Bachman, L. F. (2005b, July). *What are we assessing? The dialectic of constructs and contexts in language assessment*. Plenary presentation at the Language Testing Research Colloquium, Ottawa, Ontario, Canada.

Bachman, L., Davidson, F., Ryan, K., & Choi, I. (1995). *An investigation into the comparability of two tests of English as a foreign language: The Cambridge-TOEFL comparability study.* Cambridge, England: Cambridge University Press.

Carroll, J. B. (1961). Fundamental considerations in testing for English language proficiency of foreign students. In *Testing the English proficiency of foreign students* (pp. 30–40). Washington, DC: Center for Applied Linguistics.

Cooper, R. L. (1968). An elaborated language testing model. *Language Learning, 7,* 57–72.

Davies, A. (1997). Introduction: The limits of ethics in language testing. *Language Testing, 14,* 235–241.

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.

Hamp-Lyons, L. (1997a). Ethics in language testing. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education* (Vol. 7: Language testing and assessment (pp. 323–334). Dordrecht, Netherlands: Kluwer.

Downloaded by [National Institute of Education] at 22:54 02 May 2014

Hamp-Lyons, L. (1997b). Washback, impact and validity: Ethical concerns. *Language Testing, 14,* 295–303.
Hymes, D. (1974). *Foundations in sociolinguistics: An ethnographic approach*. Philadelphia: University of Pennsylvania Press.
Kunnan, A. J. (1995). *Test taker characteristics and test performance: A structural equation modelling study.* Cambridge, England: Cambridge University Press.
Lado, R. (1951). *English language tests for foreign students*. Ann Arbor, MI: George Wahr.
Roach, J. (1971). *Public examinations in England: 1850–1900.* Cambridge, England: Cambridge University Press.
Shohamy, E. (1993). *The power of tests: The impact of language tests on teaching and learning*. Washington, DC: National Foreign Language Center.
Shohamy, E. (1997). Testing methods, testing consequences: Are they ethical? Are they fair? *Language Testing, 14,* 340–349.
Shohamy, E. (2001). *The power of tests: A critical perspective of the uses of language tests*. London: Longman.
Shohamy, E. (2006). *Language policy: Hidden agendas and new approaches.* New York: Routledge.
Stevenson, D. P. (1985). Authenticity, validity, and a tea party. *Language Testing, 2,* 41–47.
Weir, C., & Milanovic, M. (Eds.). (2003). *Continuity and innovation: Revising the Cambridge Proficiency in English Examination, 1913–2002*. Cambridge, England: Cambridge University Press.

Downloaded by [National Institute of Education] at 22:54 02 May 2014

APPENDIX
Examinations in England in the Victorian era

In the Victorian era, under the influence of utilitarian thinkers such as Jeremy Bentham (*Constitutional Code*, 1827) and John Stuart Mill (*On Liberty,* 1859), the idea took root that it was the responsibility of government to stimulate society to improve itself. Bentham worked out an elaborate examination system for applicants to the civil service, and Mill proposed a system of compulsory education based on an examination system. Thomas Babington Macaulay, who entered Parliament in 1830, also became a noted supporter of the use of examinations for entry to public office. He developed the argument that early promise in youth is a good predictor of people's potential for future tasks and that examinations were the best way to measure this potential. Competition for entry was seen as the way to improve the quality of civil servants, and in 1853 a framework was described for making examinations the way to recruit people to the civil service. The India bill introduced competitive exams for the Indian civil service. In the field of education, exams for teachers were seen as the means of improving the quality of teachers and their teaching, and from 1846, common exams were set in all colleges. In this respect, Roach (1971) pointed out that the examination system for teachers was the first common test in England set on a general syllabus and taken in a number of separate places. The examination boards of Oxford and Cambridge had their origins around the same time, in the 1850s. In June 1857 the University of Oxford Delegacy of Local Examinations was established by statute. Its aim was to conduct exams of nonmembers of the university as part of a movement to reform universi-

ties and make them more socially involved. In spring 1857, Cambridge University received a request from schools in Birmingham, Cheltenham, Leeds, and Liverpool to offer local exams. The Council of Senate recommended that a syndicate be set up, and UCLES was eventually established in February 1858. The first examinations were held in December that year.

Downloaded by [National Institute of Education] at 22:54 02 May 2014