ıler, S. P. 1973. *Introducing ılied Linguistics.* mondsworth: Penguin.

Els, T., Bongaerts, T, Extra, G., Os, C. and Janssen-van Dieten, 984. *Applied Linguistics and the ·ning and Teaching of Foreign ıguages.* London: Edward Arnold.

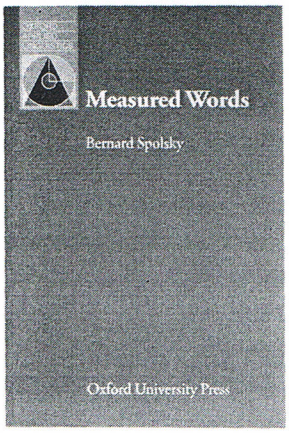
Halliday, M., Strevens, P. and McIntosh, A. 1964. *The Linguistic Sciences and language Teaching.* London: Longman.

Stern, H. 1983. *Fundamental Concepts of Language Teaching.* Oxford: Oxford University Press.

MEASURED WORDS

Bernard Spolsky

Oxford: OUP, 1995



'Ease triumphing over principle'

Since my first reading of Bernard Spolsky's masterly analysis of the history of language testing from a socio-political and economic perspective as much as a purely historical one, this simple phrase (p36) has remained lodged in my memory.

Measured Words is, by intention, no more than a part version of the whole subject. But it is told with a breadth of research detail, and a complete grasp of the subject that demands of its readers a respectful attention and reconsideration of their attitudes towards their role in the testing phenomenon.

Professor Spolsky distinguishes between test users and test writers, although the same individual may well play both roles in a specific set of circumstances. And he describes with great clarity the essential opposition between these roles. As he demonstrates, test users prefer tests to make binary decisions. It is much more useful for them to be able say 'this candidate passes' or 'that candidate fails' than to say 'this candidate knows so much about the subject'; 'that candidate can demonstrate so much ability in the area'. These are the kind of statements which are of greater interest to the test writers.

Most of *Measured Words* has to do with the design and application of tests and exams on a large scale. It describes the realities of a world in which examinations are used as a means of perpetuating meritocratic elites within societies, and in which the conflict between privilege and patronage on the one hand —the predecessors of formalised examination systems— and objectivity on the other —which is often just a mask for those same predecessors— is still unresolved.

But equally, the statements he makes, and those on which he draws from his peers and predecessors, have their value when applied to the smaller scale reality of the individual classroom. Spolsky quotes Latham who, as long ago as 1877, asked 'what we want to effect by testing and how far we can succeed' (p7); and Lowell, who in 1926, responded saying that examinations are used 'to measure progress, to control instruction, and to set a standard' (p38). This question and answer routine needs to be repeated on a regular basis by all practising classroom teachers. Seventy years after Lowell's reply, their response will probably still coincide with his.

If I consider the sheer volume of testing I and my colleagues are involved in within the Spanish educational system then the significance of this question and answer takes on mammoth proportions. The power of the

teacher is in the final exam. Through this —like 'the finger of God'— we wield life-threatening power over the individuals in our charge. The power of the examination, and the essential power relationship which it enshrouds, is a phenomenon few teachers are consciously aware of. And yet, even fewer are unaccustomed to wielding this power.

On the small scale, the tension which exists between 'ease' and 'principle' is even more real in the everyday world of the classroom. How often do we use discrete item tests —technically reliable, but maybe neither valid nor easy to interpret— instead of 'authentic', integrative test mechanisms - technically reliable but more costly in terms of both preparation time and marking time.

And on the larger scale, as Spolsky describes, there are very real social, political and economic influences on the testing instruments chosen. To what do we owe the high degree of dependence of the Spanish *reforma* on a textbook based curriculum? Similarly, the acceptance of a University entrance examination which involves the payment of a registration fee is surely questionable. To what extent is it justifiable to charge a fee when nothing has been invested in the design, trialling and administration, or post-test analysis of that test? *Measured Words* points to consequences which can arise out of a pragmatic, cost-effective approach to testing. It also reveals the impunity with which establishments that 'test' are able to continue to hold the reins.

A history in two parts

In *Measured Words,* Professor Spolsky begins by telling the story of the development of psychometric testing and its application in the field of language testing from the end of the nineteenth century. He documents the work of the British pioneer in psychometrics, Edgeworth (1888) and follows the many lines of research through the years of both World Wars. He then diverges from the broad picture of testing to focus in the second part on the development of the American TOEFL test, as it is known today

Underlying the two part structure of *Measured Words* is the contrast Spolsky makes between two intellectual approaches: the Cartesian, rationalist, and the skeptical humanist. The first, he classifies as representing an empiricist attention to measurement which he describes throughout the book as the hunt for technical reliability; and the second, as a descriptive approach, which he attaches to the less scientific practice of the search for validity through a meeting of erudite minds. The former, is identified loosely with the American TOEFL examination and the academic, investigative background baggage attached; the latter, with the traditional British university 'wise men' approach, which is linked with the University of Cambridge (UCLES) examinations. *Measured Words* concentrates on TOEFL, that is on the objective, scientifically oriented approach. But Professor Spolsky makes it clear throughout that whichever path you follow, both are part of an institutional-industrial approach to testing that, in the real world, generates enormous amounts of money.

The decision to follow the development of TOEFL and not the British-based Cambridge examinations —which are more familiar to us in Spain— is clearly justified in the course of the book: TOEFL is the perfect example of the kind of conflict of interests which Spolsky has found throughout the development of an increasingly commercial testing industry. But at the same time it represents the culmination of a thoroughly scientific concern for the justification of decisions taken when examining individuals. The striving for technical accuracy in the case of TEOFL is far older and more transparent than it has ever been at UCLES.

By contrast, the author's brief description of the Cambridge examination reveals the lack of the concern, up until relatively recently, for this type of investigative interest. Spolsky's primary interest in describing the empiric basis on which TOEFL examination judgements are made, was not shared in the UK.

From a historical perspective, then, Spolsky views the political, social and economic constraints on testers which greatly influence the development of their tests. These real world pressures, he shows, compete with the theoretical input from language learning and psychometric research, and often lead test writers and tests to the 'quick'n'dirty' conclusions we see around us.

Professor Spolsky looks at three periods within the history of testing which he calls the Traditional, the Modern, and the Post-Modern. He demonstrates the length that history has by

describing the work of Edgeworth, dating from the end of the nineteenth century, right up to the present day. This is the list of researchers which Spolsky cites as essential background reading for anyone interested in Testing.

The reliability versus validity debate on which Spolsky focusses is one he develops throughout the book. In the chapter devoted to the historical development of prognosis, or aptitude testing Spolsky perceives two distinct periods: one before, and one after the work of J.B. Carroll. Prior to Carroll Foreign language learning had been based on the Grammar-translation approach, and as such had been viewed as reasonably successful. Aptitude testing had been used as a means of selecting those learners who would follow courses in modern languages and its purpose principally was 'to keep prospective failures out of classes' (p117): 'what Cheydleur (1932, quoted by Spolsky p117) called the 'mortality of modern languages students'.

The theoretical issues which arose at that time were those to do with the nature of language ability and, most specifically, whether or not it could be viewed as being unitary or divisible into identifiable, and therefore testable, separate abilities. Carroll's most recent work (1993), quoted by Spolsky indicates that an answer has yet to be found.

Spolsky's research documents in particular aptitude tests produced over the 1920s and 30s. And in his analysis he touches on these prime issues. Spolsky asks one fundamental question over and over again: we have many many tests, but what do they measure? The theoretical argument he exposes is that within the ranks of testers there is much concern for the reliability of tests. They want to know whether one specific test, taken by a group of learners will produce the same results each time. However, once that kind of reliability has been statistically established we need to ask ourselves as to the validity of the test. The more important question is not 'Are we doing a statistically good job?' but 'Are we doing the right job?'

The fundamental question in testing must surely be that which underlies the purpose of the test. Spolsky charts the historical use of tests as a means of establishing and maintaining a power relationship. He makes a parallel between the pedagogical test and the speech act. A speech act both informs and performs: it transmits information between the interlocutors and it performs in that it establishes the nature of the relationship between them *viz a viz* that information. The realities of this metaphor can be seen easily if we transpose Spolsky's argument to the public sector in Spain. Here we see 'qualifying tests' *(oposiciones)* everywhere. They are restrictive entry tests used to limit access to social classes. As such, they perform a strict controlling role on Spanish society.

In schools and universitites we also use tests and exams as a means of controlling the educational and social progress of individuals. The system is actually moulded to benefit the economically better-off classes - they are the students who can afford to repeat courses until they pass them, thus assuring their ultimate access to graduate status. The test mechanisms by which we exert this power are probably neither reliable nor valid in objective empiric terms. They are a part of what Foucoult described as 'the discipline of education' (p34). And yet teachers 'reify' results. The magic power of '5 = a pass' is something so deeply ingrained in teachers and students as to be totally beyond belief. '5' represents success at overcoming a hurdle, an obstacle in the race. But in terms of real world knowledge and learning and ability to perform, what does it mean? The limitations of testing were identified by Edgeworth and since then the wheel has been reinvented *ad nauseam.*

Is Professor Spolsky's overview of testing essentially optimistic or pessimistic? I find it hard not to feel somewhat disheartened, at the least, on a full reading. Much has been done in the field of reliability and little in that of validity. Much has been done to produce 'effective' binary tests and little to create mechanisms which really test what we would claim to be teaching. Rightly, he points to the significant contribution made by Lyle Bachman in offering a theoretical model to underpin test development; rightly, too, he applauds the efforts made by TOEFL and UCLES to research and adapt to the challenge of communicative English language testing for the next century. But the question of what valid and reliable communicative language tests may look like is left for that future.

Bryan Robinson