DOCUMENT RESUME

ED 340 207 FL 019 888

AUTHOR Spolsky, Bernard
TITLE Of English Marks and American Reviewers.
PUB DATE Mar 90
NOTE 15p.; Paper presented at the Annual Meeting of the Teachers of English to Speakers of Other Languages (24th, San Francisco, CA, March 6-10, 1990).
PUB TYPE Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; *Cultural Differences; *English (Second Language); Foreign Countries; *Interrater Reliability; *Language Tests; Scores; *Scoring
IDENTIFIERS Americans (United States); British; *Cambridge English Examinations; *Test of English as a Foreign Language

ABSTRACT

A discussion of the differences between the Test of English as a Foreign Language (TOEFL), an American test battery, and the Cambridge English Examinations (Cambridge), a British battery, focuses on the different approaches to language test development embodied in the tests as the source of difficulty in translating between them for individual examinees. Drawing on the results of one statistical comparative study, two alternative approaches to test comparison are suggested: (1) considering the effectiveness of the two test batteries in achieving their tasks; and (2) examining their usefulness for making admission, proficiency, and instructional decisions. The development of the two tests is viewed from both philosophical and historical perspectives. It is suggested that the tests emerged from different world views, one emphasizing a scientific approach and characterized by unease with personal decision-making, and the other a traditional humanistic approach characterized by distrust of formulas and mathematical predictions. A 17-item bibliography is included. (MSE)

Paper prepared for the Colloquium on the Cambridge-TOEFL Comparability Study, TESOL Convention, San Francisco, March 1990.

Of English Marks and American Reviewers

Bernard Spolsky

Bar-Ilan University

Philosophies of testing

The study by Bachman, Davidson, Ryan and Choi (1989) set out to perform its difficult task of comparing TOEFL and the equivalent Cambridge tests by direct statistical and content comparison of two sets of tests.(1) It will be generally agreed that the study was carried out meticulously and thoroughly; within institutional, methodological and practical limitations that are clearly and exhaustively described, the tests were administered to a large and varied sample of candidates, and various interesting statistical analyses of the results have been reported. While there is room for discussion of the appropriateness of some of the statistical models and for much fruitful debate over their interpretations,(2) the basic finding of the comparability study has been clearly presented: it found good evidence of overlap between the two test batteries in the abilities they measure, but had difficulty in establishing any simple way of translating between them for individual candidates.

There are some clues in the test content analyses as to where some of the causes of these differences might reside. A more fundamental explanation of difference is offered in the suggestion that the two test batteries "represent radically different approaches to language test development," with the TOEFL seen as representing the "prototypical" psychometric- structuralist language test (Spolsky 1977), and the CPE and FCE as forms of traditional examinations, with the mediation of an authorized judge moderated by the control of an experienced chief examiner. Given such a basic difference, it was perhaps inevitable that the results of the statistical analyses should come out as they did, making uncomfortably clear the failure of the Cambridge tests to reach the standards of reliability that they have, quite openly, been slow in

accepting as basic or necessary. There is already evidence that the review process will lead to a number of changes in the procedures fo r administering and scoring CPE and FCE. It is perhaps a cause for regret that the TOEFL has not been submitted to the same kind of scrutiny from an outside standpoint, for Bachman and his colleagues, by choosing to stand mainly on psychometric ground, have no other criteria available to make their judgments by. One promising attempt to find a neutral stance, the test content analysis, unfortunately failed to capture the "qualities that may be valued by the British measurement tradition."

With hindsight made feasible by the pioneering work of the study, it is possible to imagine alternative approaches that might have been tried. One such strategy might have been to consider the functional effectiveness of the two sets of tests: how well do they achieve their tasks? We might have been presented then with evidence on at least two relevant questions: how helpful are the two test batteries in making admission and other proficiency decisions, and what useful or harmful effects do they have on English language instruction?

The first of these questions is obviously a basic one, but it is interesting to note that here too there is a philosophical difference in the approaches of the two testing agencies. ETS assumes that tests can be used to produce numerical scores which rank candidates on a normal curve on some relevant ability or combination of abilities and that subsequently, these scores will be used to make some locally appropriate decisions. Following this, a test is satisfactory (to the test makers) to the extent that it provides statistically tidy rankings, and (to the test users) to the extent that the interpretations they make of the scores are accepted by candidates. The Cambridge Syndics assume that an examination can determine who has reached or failed to reach an intuitively known standard; that examinations in other words will cluster candidates appropriately for them to be labelled outstanding, good, satisfactory, or failures. factory (to the test makers) if it enables them to make these judgments and (to the test users) if the candidates and their future and past teachers accept these results.

The difference in fact is wider. In his consideration of the report by Bachman and others, Peter Strevens suggested that the two test batteries might be considered as the products of two different

paradigms:

> One of these sees the task as properly the domain of psychometric measurement, that being defined in ways that make the prior educational experience of the candidate irrelevant, while insisting on the utmost attention to validity, reliability, and statistical processes. The other paradigm views the assessment of EFL performance as a part of the total educational process, and inseparable from it, to the extent that expert judgments are admissible in lieu of statistical data, if the total educational process so require. (Strevens, 1989:1).

In spite of this dissimilarity in philosophy or paradigm, it is theoretically possible to investigate how successful the tests are on these more general criteria, to check in other words how often candidates and those who know them (before and after the test) consider them satisfactory measures of the desired abilities. Detailed validity studies of this kind have been reported from time to time for TOEFL and for ELTS, the equivalent English proficiency testing batter, also administered by UCLES, but to carry them out is difficult.

The effect on instruction is another basic question on which the comparability study gives us no information. The Great Divide between supporters and opponents of objective tests is marked essentially by two different answers to the question of the effect of objective tests on instruction: the supporters believe that it has no deleterious effect, and that symbolic use of context (in, for instance, the TOEFL vocabulary items, where it remains for its washback (3) effect on instruction rather than for any recognition of its value in the item) is enough; the opponents believe that multiple choice testing can only lead to sterile teaching. Given the critical importance of this argument in the common rhetoric, it would be valuable to have some empirical evidence one way or the other.

A quite different approach might have been to consider not the effectiveness and effect of the two tests, but how they compare to some general notion of an ideal language test. One way to do this would be to take more seriously the recent attempts to describe the complexity of language proficiency such as Spolsky (1989a, 1989b) or Bachman (1990) and to ask which aspects are more relevant to the tasks set for

these tests. Such an approach would have made clearer the extraordinary rigidity and narrowness of focus of TOEFL (as of most psychometrically satisfactory tests), a point well made by the general insistence on unidimensionality, and recognized in recent years by ETS in its finally meeting the demand (first expressed in 1961) for integrative tests, and in particular for direct testing of speaking and writing. Another way of getting at this might have been to note the way that both the tests have in fact been moving over recent years, ETS to deal with the limitations imposed by its deification of reliability, and Cambridge to recognize the need to make explicit and reliable the judgments it treasures.

A historical view might perhaps allow a wider perspective on these issues.

A historical view

The comparability study (Bachman and others 1989) has taken, not at all unreasonably, a synchronic view of its task: its mission was to compare the present form of TOEFL and the UCLES tests. Encouraged by impressions I developed at the Advisory Committee meeting last year of a seeming philosophical or ideological gap between the two institutions, I have started to gather data to support or modify the notions of language test history that I first proposed some years ago (Spolsky 1977). A synchronic analysis makes clear why there is so much scope for the tests to vary: the enormous complexity of language proficiency (Spolsky 1989a,b; Bachman, 1990) and the physical, temporal and fiscal limitations on any real-life practical testing mean that any two tests, however similar in approach and goal, may easily and reasonably choose quite different aspects to measure and quite different ways of measuring. A diachronic study should help understand why a test has chosen the features it has, and even more, to help appreciate whether two contrasting tests should be seen as converging or diverging.

While my 1977 paper shows that I am not particularly fearful of making large claims before collecting too much data, now that I am actually engaged in the historical study, I am newly reluctant to rush to conclusions. Here, then, I would like to present some very tentative notes, with hints of the data that might make them more conclusive.

One of the reasons for my continuing fascination with language testing is the fact that it constantly sets practical and theoretical issues into fruitful tension: the needs of the tester regularly challenge the theorist, just as the findings of the theorist repeatedly tempt the tester. While it is fairly easy to come up with new assessment procedures, it remains difficult to explain exactly what is being measured, a situation that guarantees a continuing productive stress. As if this first cause of tension were not enough, there is a second one provided by the fact that at least two disciplines claim proprietary rights in the theory behind language testing: both language learning theorists and measurement experts have their own independent notions of what is involved. If I might reinterpret the earlier suggestion I made about the three periods of language testing (Spolsky 1977), one might consider the first as a period when only language learning theory was considered relevant, when there was in fact no measurement theory readily available or willingly applied; the second might be characterized as a period (or approach) that assumed that all one needed to do was to add together two independent theories, structural linguistics which mandated the items to be measured and psychometrics which would determine how to measure them; and the third period (the post- modern), a time of recognition that what is required is a combined and synthesized approach to language assessment, one that can only come from open and receptive intercourse between linguists and psychometrists.

It is only a slight exaggeration to consider many public examinations in Britain and other parts of the world as still dominated by the first approach, with their continuing suspicion of the objective test and their cautious incorporation of the psychometric theory that goes with it; similarly, the second approach governs the purely objective tests of many US and US- influenced testing institutions. The third approach is already adumbrated by the tensions in each of the other two, as witness the inclusion of objective sections in British examinations, and the reluctant steps finally taken to include direct measures of writing and oral ability in US tests.

The power of these competing views helps explain how the conference in 1961 that led to the development of the Test of English as a Foreign Language (Anonymous, 1961) managed to resist arguments for an impressionistically marked composition and to postpone until after further research a

test of oral production, gaps that some quarter of a century later have been tentatively filled, this in spite of the fact that the major theoretical paper at the conference was a call for integrative testing (Spolsky 1990a). Similarly, although the English Foreign Language tests of the University of Cambridge Local Examinations Syndicate (1987) have succeeded in keeping their main emphasis on "marks gained in more traditional ways, i.e., those awarded on impression for performance in various communicative tasks," they have, since the late 1960s, included alternative, and since the late 19?0s, compulsory objective Papers which are used to cross- reference and adjust the subjective marks. In this way, two tests starting out from contrasting philosophical views of the nature of testing are slowing moving to a principled eclecticism that combines the two positions. Bachman and others' report would be strengthened, I suspect, by recognizing this: by noting the American concern for developing direct (and thus psychometrically less rigidly controlled) measures as a converging tendency alongside the British emerging recognition of the value of reliability.

It is intriguing to speculate on how the two groups of testers managed, thirty years ago, to seize and hold on to quite different part of the testing elephant. Reading the report of the 1961 Conference that led to the development of TOEFL (Anon. 1961), there are no hints of any questioning of the fundamental principles of objective testing, except in some references in a report on British Commonwealth testing (MacKenzie 1961). None of the participants whose accounts of the meeting I have so far been given have any recollection of such a discussion, with two exceptions. The first concerned oral production, which the conference agreed would need to be delayed until research could develop appropriate (i.e., objective) techniques, a task left for twenty year s or so. This, it must be noted, happened in spite of the fact that one local (4) testing program, the American University Language Center program, already made use of an oral interview not unlike the Foreign Service Institute Oral Interview (Harris,1961)) The second concerned writing, but reports of the discussion suggest that persons who spoke for including a direct writing test (5) were convinced by the testing experts such as John Carroll and Fred Godshalk of the "great difficulties (both technical a nd logistic) entailed in any professional scoring of such samples" (Carroll (1989 private communication). The conference agreed then to include an objective writing test of the kind developed by Godshalk and others for the College Entrance Examinations Board's tests, and to

collect a writing sample to be forwarded to university's to use as they wished. In fact, this was not done, but once again, twenty years later, a writing test has now been developed and made available.

There were about 20 people present at the Washington meeting. A number of them had professional training as language testers: Carroll himself, and Robert Lado (whose book on language testing had just appeared, published first in England); Godshalk of ETS; David Harris and Les Palmer (the first and second directors respectively of the TOEFL program); Sydney Sako (trained in educational psychology at Texas and testing director at Lackland); clearly, this group set the tone for the basic acceptance of psychometric principles for objective testing in the "fundamental considerations" that went into test design. It is not surprising then to find only tentative arguments put forward for traditional testing. What is to be noted however is the recognition in a number of ways of the needs of the post- modern approach: the call for research into techniques for testing oral production (which, of course, was going on at that very time on the other side of Washington at the Foreign Service Institute; (6) the concern for the writing test; the acceptance of Carroll's proposal to include a language aptitude test.(7)

The Washington meeting provides a fortunate body of data for the test historian.(8) My study of the British developments at the same period are at a much earlier stage, and I can simply mention some guesses based on hints that I have received. In 1960, the Cambridge tests were in a period of rapid growth: there were by then over 25,000 candidate annually, and the system of traditional scoring was well in place. The people in charge of the tests were none of them trained in psychometrics, but seem to have come to their position from a modern language background. J.O. Roach, for instance, who was deputy secretary of the Syndicate from 1925 to 1945 and played a major part in creating the Cambridge examinations in English for foreign students was a lecturer in modern languages at the University. In the late 1930s and into the 1940s, the questions raised by Hartog and Rhodes led to some studies of reliability (Spolsky, 1990b), but this did not lead to changes in test form. There are hints that in the early 1960s UCLES learned about the values of objective testing, from both native British and visiting American experts.

The pressures of this knowledge led first of all to the addition, in the mid-1960s, of semi-objective alternative papers testing usage and vocabulary; these alternative papers were fully incorporated in the examinations in the 1975 integrated syllabus. We might reasonably ask why the influence of British language testing experts (one thinks of Peter Strevens whose paper on objective testing was written for a 1960 Makarere meeting and circulated at the 1961 Washington meeting, and Alan Davies whose 1968 book provides such a sound review of language testing) was so slow to be felt.

Strevens (1989) set out to explain some of the background. He pointed out that the Cambridge examinations were developed largely under the influence of the staff of the Department of English of a Foreign Language of the University of London Institute of Education or British Council officers with training in that department. This department was then training most of the people working in developing "belatedly" English language education in countries of the British Commonwealth. While it followed methodologically in the tradition of Jespersen, Daniel Jones, and Palmer, a "modified oral approach," there were no applied linguists in the department. The dominant British education model at the time was the primary school, and successful primary methods, with notions of a "happy classroom" and informal ongoing assessment, were developed for secondary schools too. Formal assessment came only at the end of secondary school, as part of examining a candidate's performance in all subjects learned, as part of accreditation for a job or (occasionally) for further education.

Strevens himself came into contact with the Cambridge examination when, after having served on the staff of the School of Applied Linguistics at Edinburgh, and having been appointed to a new Chair of Contemporary English Language at Leeds, he was invited to serve on the UCLES EFL executive committee. I quote his words on the experience:

> I had acquired a reputation for saying publicly that the UCLES exams, FCE and CPE, were already old-fashioned and should be modernized.... After two years with no visible signs of change in the exams being accepted, I rebelled, at a stormy meeting where the Chairman, with the full approval of the committee, said to me that the reason why the Cambridge exams should not be changed was that "...they force the teacher to teach according to the best possible methods."

(Strevens, 1989:5)

The examinations, Stevens argued, were intended to control the instruction process rather than to assess proficiency. They were, he was in 1989 ready to concede, "exercising responsibility for bringing about changes in the teaching syllabus and in classroom methodology," while resisting any changes to the examination.

My own studies to date suggest the social and institutional foundation of change in testing technology and practice. There was not in fact a scientific or theoretical wall blocking intercourse across the Atlantic. The development of US psychometrics owes much to the work of Hartog and Rhodes, and there were plenty of British scholars who were aware of and who have contributed to the development of testing theory. For instance, the study by J.P. Roach (1945) carried out for the Cambridge Syndicate shows a very sophisticated understanding of reliability and validity problems with oral testing, a topic not treated in detail in the language testing literature for another twenty-five years. But our concern here is with institutional tests, something that by definition are slow to change. Once TOEFL was set in its first form (and this first form reflected the current practice of existin g US tests for foreign students as well as the current views of language testing of the experts present at the meeting), and especially once the test moved to ETS and was locked into the complex production and administrative structure required for such an industry, modification was slow. By 1960, the Cambridge tests too were already firmly locked into their own institutional mold. In each case, it is a mark of the strength of ideas that modifications could even be considered.

Two world views

Underlying the difference of approach of the two testing bodies, and going even beyond the issue of the effect on teaching of an examination, is in fact a difference of more general interest, the gap between humanists on the one hand, with their distrust for formulas and mathematical predictions, and the scientists, with their unease in the face of unexplained personal decision-making. The same controversy underlies the differing approaches taken by consultants to decision-making. As a recent review arti cle by

Dawes, Faust and Meehl (1989) presents the issue, when experts (physicians, psychiatrists, psychologists) are consulted on individual cases and asked to diagnose cases or made predictions about outcomes, they have two ways they may choose to interpret whatever data they collect: either they use clinical methods or actuarial methods. This distinction refers not to the collection of data (which may well be clinical) but to the decision making. Dawes, Faust and Meehl categorize the two methods as follows:

> In the clinical method the decision-maker combines or processes information in his or her head. In the actuarial or statistical method the human judge is eliminated and conclusions rest solely on empirically established relations between data and the condition or event of interest. (1668)

I am confident that the analogy to our question here will be fairly obvious: in what I have called the traditional or pre-scientific approach, the decision or judgment is left mainly to the human examiner; in the psychometric or modern approach, the human examiner is eliminated, and scores are established on the basis of statistically validated items and tests. Over the last thirty years or so, Dawes, Faust and Meehl report, there have been close to a hundred studies comparing the accuracy of the clinical and the actuarial approach, and they basically agree that the odds clearly favor the actuarial: even when conditions of the study seem to favor the clinical judgment (some extra information, or access to the actuarial results as well), the actuarial method still surpasses the clinical judge in accuracy. Most of the studies cited are in the medical or psychiatric fields, but included are ones closer to our interest such as prediction of success of university students.

Dawes, Faust and Meehl discuss why the actuarial or statistical methods are superior. First, they are consistent in decisions, and do not suffer from the random fluctuations of human judgments. More important, when they are based on correct models, they are consistent in their weighting of the variables, making sure that each contributes its established predictive power to the decision. For a number of reasons, human judges are inconsistent in their weighting of variables, often being influenced by skewed experience or overconfidence.(9) Having made the case for actuarial or statistical methods, Dawes, Faust and Meehl conclude by drawing attention not just to benefits but also to limitations. Clearly, the methods need to be based on sound models and good empirical studies, and this is not always the case. One of the

most important benefits of the approach is that it makes the basis of decision-making explicit and thus open to criticism and modification. Finally, they point out, actuarial methods

> reveal the upper bounds in our current capacities to predict human behavior. An awareness of the modest results that are often achieved by the best available methods can help to counter unrealistic faith in our predictive powers and our understanding of human behavior. It may well be worth exchanging inflated beliefs for an unsettling sobriety, if the result is an openness to new approaches and variables that ultimately increase our explanatory and predictive powers. (Dawes, Faust and Meehl 1989:1673)

If nothing else, we can agree that the comparability study has surely contributed to the ongoing dialogue that guarantees the questioning of strongly held beliefs in the perfection of present approaches.

Notes

(1) This paper, prepared for the Colloquium on the Cambridge- TOEFL Comparability Study, TESOL Convention, San Francisco, March 1990, is dedicated to the memory of Peter Strevens.

(2) Henning (letter in press, *Language Testing*, raises a number of these issues. Alastair Pollitt has suggested a different approach to use Rasch analysis for comparison.

(3) For discussion of washback in testing, see Hughes (1989).

(4) For students of transatlantic English, one might point out that the term "local" here means the testing batteries of individual US universities. The term "local" in the name University of Cambridge Local Examinations Syndicate means that while the tests are prepared (and may be scored) at Cambridge, the candidates may take them locally, in their own countries.

(5) It seems to have been Joel Slocum, an admissions officer, who argued this case most strongly.

(6) The first published account of the FSI test was in a report by Frank Rice published in the *Linguistic Reporter* in May 1959. The first scholarly analysis did not appear until 1975 (Wilds 1975).

(7) This was agreed to by the Conference but never carried out.

(8) See Spolsky (1990)

(9) Roach's discussion of how examiners arrive at their marks shows full appreciation of this fact.

References

Anon. 1961. *Testing the English proficiency of foreign Students. Report of a conference sponsored by the Center for Applied Linguistics in cooperation with the Institute of International Education and the National Association of Foreign Student Advisers.* Washington, D.C.: Center for Applied Linguistics of the Modern Language Association of America.

Bachman, Lyle. 1990. *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L., Davidson, F., Ryan K., and Choi, I-C. 1989. *An investigation into the comparability of two tests of English as foreign language: The Cambridge-TOEFL comparability study.*

Dawes, R.M., Faust, D., and Meehl, P. 1989. Clinical versus actuarial judgement. *Science,* 243:1668-74.

Harris, David P. 1961. The American University Language Center Testing Program. In Anonymous (1961), 41-53.

Hughes, Arthur. 1989. *Language testing for teachers.* Cambridge: Cambridge University Press.

MacKenzie, Norman. 1961. English proficiency testing in the British Commonwealth. In Anonymous (1961), 54-72.

Rice, Frank. 1959. The Foreign Service Institute tests language proficiency. *Linguistic Reporter,* (May 1959).

Roach, J.O. 1945. Some problems of oral examinations in modern languages: an experimental approach based on the Cambridge Examinations in English for foreign students, being a report circulated to oral examiners and local examiners for those examinations. Cambridge: Local Examinations Syndicate.

Spolsky, B. 1977. 'Language testing: art or science' in Gerhardt Nickel (ed.): *Proceedings of the*

Fourth International Congress of Applied Linguistics. Stuttgart, Hochschulverlag. Volume 3, pages 7-28.

Spolsky, Bernard. 1989a. 'Communicative competence, language proficiency, and beyond.' *Applied Linguistics* 10:138-56.

Spolsky, Bernard. 1989b. *Conditions for second language learning*. Oxford: Oxford University Press.

Spolsky, Bernard. 1990a. The prehistory of TOEFL. Paper presented at the Language Testing Research Colloquium, San Francisco, March 1990.

Spolsky, Bernard. 1990b. Oral examinations; an historical note. Paper read at the 1990 annual meeting of the Israeli Academic Committee on Research in Language Testing, Kiryat Anavim, May 1990.

Strevens, Peter. 1989. Comments on Bachman and others. (MS)

University of Cambridge Local Examinations Syndicate. 1987. *English as a Foreign Language: General Handbook*. Cambridge: University of Cambridge Local Examinations Syndicate.

Wilds, Claudia. 1975. The Oral Interview Test. In Randall Jones and Bernard Spolsky (edd.) *Testing Language Proficiency*. Washington: Center for Applied Linguistics.