Annual Review of Applied Linguistics (2009) **29**, vii–xii. Printed in the USA.
Copyright © 2009 Cambridge University Press 0267-1905/09 $16.00
doi:10.1017/S0267190509090011

EDITOR'S INTRODUCTION

Bernard Spolsky

From the beginning, public tests and examinations were instruments of policy. The Imperial Chinese examination was created to permit the emperor to replace the patronage system by which powerful lords were choosing their own candidates to be mandarins. The Jesuit schools in 17th-century France introduced a weekly testing system to allow central control of classroom teaching. In 19th-century England, Thomas Macaulay argued for employing the Chinese principle in selecting cadets for the Indian Civil Service; a similar system was later used for the British Civil Service. A primary school examination system was set up in England at the end of the 19th century to serve the same purpose of achieving quality control and accountability in public schools as was proposed for the No Child Left Behind Act (NCLB) that is being bitterly disputed in 21st-century United States. Chauncey's primary goal after World War II in developing the Scholastic Achievement Test for admission to elite U.S. universities was to replace the children of the wealthy establishment with highly qualified students who would see their role as contributing to public service.

Language tests also served political functions. The English tests for foreign students first created by the U.S. College Board in the 1920s, tried again immediately after the World War II, and finally enshrined in the Test of English as a Foreign Language, were responses to a request from the U.S. Commissioner of Immigration for a method of blocking the loophole left in the 1927 Immigration Act allowing free admission of students coming to U.S. schools and colleges (Spolsky, 1995). Even more striking was the use by the Australian Immigration Service from the early 20th century of a dictation test to block undesirable immigrants (Davies, 1997; McNamara, 2005); language tests for immigration and citizenship are becoming increasingly common.

When I started my own involvement with English language testing, I quickly became aware of the political significance of requiring English knowledge as a major criterion for admitting foreign students to American universities, realizing the way

https://doi.org/10.1017/S0267190509090011 Published online by Cambridge University Press

that it guaranteed places to the children of wealthy members of the establishment (Spolsky, 1967). In more recent years, there has started to be widespread appreciation of the intimate connections between language testing and language policy, led by the clarion call of Shohamy (1993). I greatly appreciate the decision of the editor-in-chief and the editorial directors of the *Annual Review of Applied Linguistics* to devote this issue to the topic.

Over the last 5 years, assessment has been treated occasionally in ARAL:

1. 2007: an article by Dan Douglas and Vogel Hegelheimer, "Assessing Language Using Computer Technology."
2. 2006: article by Cathie Elder and Alan Davies, "Assessing English as a Lingua Franca."
3. 2005: (in a survey volume): section on research in language assessment with three articles: Thom Hudson, "Trends in Assessment Scales and Criterion-Referenced Language Assessment;" Joan Jamieson, "Trends in Computer-Based Language Assessment;" and Penny McKay, "Trends in the Assessment of School-Aged Learners."
4. 2002: (in a volume on the theme of discourse and dialogue) section on discourse analysis and assessment with three articles: Tim McNamara, Kathryn Hill, and Lynette May, "Discourse and Assessment;" Richard Young, "Discourse Approaches to Oral Language Assessment;" and Ulla Connor and Aymerou Mbaye, "Discourse Approaches to Writing Assessment."

The recent interest in the topic is highlighted in a vigorous correspondence being conducted at the moment on the ELLADVOC list on how to influence the education policy of the new Obama administration. James Crawford, leader of the group, set out the difficulty of the task:

> Our main challenge now is figuring out how to deal with liberals and Democrats who are wedded to the NCLB-high-stakes-testing-punitive-accountability-war-on-teachers agenda. There are many of these people trying to influence the Obama administration, and they have considerable power and influence. Their ranks include the Center for American Progress (which is deeply involved in the transition), the big foundations—Gates, Aspen, Carnegie, and others—numerous members of Congress, editorial boards at major papers like the *Washington Post* and the *New York Times,* and a number of groups that should know better but remain willfully blind to the harm their policies are doing to minority kids—for example, the National Council of La Raza, MALDEF, and the Leadership Conference on Civil Rights.

The heart of the controversy, focused as I write this on the selection of a new U.S. Secretary of Education, are contrasting beliefs about the power of tests as

https://doi.org/10.1017/S0267190509090011 Published online by Cambridge University Press

methods of assuring quality education. There are progressive reformers who believe that tests not only measure but are also the best way of assuring accountability, and there is a second group of scholars who believe that standardized testing is a procrustean bed that skews and distorts education. The essays in this collection are written by language testers who would be assumed prima facie to be supporters of the industry they depend on, but who will be seen to have learned from experience and observation how dangerous are the instruments they work to perfect.

The opening essay by Glenn Fulcher deals with this fundamental issue. It is not just poor tests, or misused tests, but the very attempt to establish collective testing that leads to imposing conformity. He proposes a method by which test writers and planners can try to understand in advance the effect of the use of their test, and so recognize their responsibility for results.

Lynda Taylor deals with the problem created by lack of understanding of testing by people who are not professional testers and who ascribe automatic reliability to any test. It used to be that tests and examinations were considered errorless because of the authority of the senior professors who conducted them; with the development of the psychometric industry, many ascribe the same infallibility to the computers that score and sometimes produce them. She urges a new test wiseness, where professional testers, applied linguists, and other decision makers appreciate the difficulties and unavoidable uncertainties (Edgeworth, 1890) of testing.

It has long been believed and is still widely held that the citizen of a democratic nation must know the official language and so be able to handle civic responsibilities. In the United States, the article by Antony John Kunnan tells us that this has been required by law for over two hundred years. As he analyzes the various provisions and tests that have been used, he finds good reasons to question the methods and even the requirement itself. As mentioned earlier, policies such as the Australian use of dictation tests in a language a prospective immigrant does not know (Davies, 1997) and the growing use of tests for immigrants in Europe (a friend of mine, an Australian professor of linguistics at the University of London, was required to pass an English test to renew his work permit) have led to increasing discussion of the role of language proficiency and tests in citizenship (Milani, 2008).

One of the clearest examples of the use of language assessment as a method of pursuing centralized language education policy is provided by the work of the Council of Europe and the European Union described by Neil Jones and Nick Saville in their article. Beginning in the 1970s, various working groups of language teaching experts under the aegis of the Council set out to try and standardize and improve the teaching of foreign languages in Europe. The main focus was the development of the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). This ambitious and thorough document, representing a summary of the state of the art in language teaching, formed the basis for the development of curricula and the presentation of a standard scale for assessment. Jones and Saville describe the background and development of CEFR and its subsequent influence on language testing and teaching in Europe and beyond. Soon, it

https://doi.org/10.1017/S0267190509090011 Published online by Cambridge University Press

will be used to survey the effectiveness of language teaching in the member countries of the European Union. Nations are being encouraged to use it, as it is being used in France, as the basis for reform of foreign language teaching. It is the assessment component that adds to the CEFR the power to affect implementation of policy.

There are, however, those who agree with the long-standing complaint of Latham (1877) that using assessment simply to evaluate teaching and learning leads to narrowing of goals and loss of potential richness. Constant Leung and Catriona Scott describe the movement to restore focus to learning, with the emphasis on formative rather than summative assessment, that has been taking place in the educational systems of Wales and Scotland, especially with the greater freedom allowed since devolution. Both systems have replaced a test-dominated regime with priority for teaching and learning. They argue that a similar approach would benefit the teaching of English as an additional language in the United Kingdom.

One of the effects of the widespread diffusion of English has been the development of local national and regional varieties. Whereas once the only problem for a teacher of English as a foreign or second language was whether to teach and test British or American English, now, as Alan Davies describes in his article, there are regular debates about which varieties of World English to test. It is, he suggests, a language policy question: Although most countries accept local varieties as appropriate for vernacular speech, they are less inclined to accept them as a standard for formal tests of speaking and writing.

In his article, Alister Cumming sets out to explore the relationship among tests, curricular policies, and teaching. Which drives which, and in what direction? And how should each be determined? Cumming notes that assessment (as in citizenship tests) is independent of teaching. At the same time, criteria-referenced standards (like the CERF described by Jones and Saville) are being used to set pedagogical approaches. And demands for benchmark standards are setting problems in handling nonstandard students (e.g., immigrants). He does not offer a solution, but helps clarify the questions that must be answered to decide how tests relate to policy.

All these have reached a major focus as a result of the recent U.S. attempts at using standardized tests to guarantee accountability in education, the strangely named No Child Left Behind Act that according to critics is leading to exclusion of sectors of the school population. The underlying issues from a testing point of view are the nature of validity and whether it includes the impact of a test on the system, and the responsibility of testers for the way tests that they develop are used. The article by Kate Menken analyzes the policy underlying the NCLB testing and presents the problems it has produced for U.S. education struggling as it is to deal with the integration of new immigrant pupils and other English language learners.

Accepting this background, Micheline Chalhoub-Deville proposes the use of social impact analysis (SIA), which was developed in social science and environmental research to mitigate the harmful effects. Recognizing the power of

https://doi.org/10.1017/S0267190509090011 Published online by Cambridge University Press

tests to produce harmful as well as beneficial results, she suggests a way that test developers can recognize and fulfill their responsibilities.

Echoing Lynda Taylor's concern for test-wiseness, and the other articles that show how even highly professional testers can produce unwelcome side effects, Hossein Farhady and Hora Hedayati describe a case that is probably much more common—an educational establishment that uses tests for control without providing evidence of the reliability and validity of the instruments it administers. I suspect this story can be repeated in many countries. What makes the Iranian case more remarkable is the recognizable inclusion of political and religious criteria in admissions testing; what it shares with many other systems is the belief that testing can somehow make up for inadequate teaching and learning.

Turning from school-based testing to testing for specific business purposes, Liz Hamp-Lyons and Jane Lockwood describe the major economic importance of the outsourcing to the Philippines of call centers and the consequent problem of developing English tests that usefully assess the qualifications of employees. The obvious first solution was to try to use available tests, such as standardized tests of academic English or even commercially available tests for specific purposes. Given that these did not seem to help, and given the expense of developing small specialized tests, they propose a feasible and effective solution to the problem. They conclude with an account of policy responses to this issue by the various interested parties in the Philippines—government, business, and the academy.

In arguments about testing, it is commonly conceded that there is ample justification for diagnostic tests intended to improve teaching (although Farhady shows that simply mandating them does not guarantee implementation) and as for qualification tests—would you like to be operated on by a surgeon who didn't pass her medical exams? In the last article, J. Charles Alderson (generously agreeing late to write a article originally promised by Harold Ormsby, who was forced by illness to withdraw the offer—we wish him a continued and full recovery) traces the history and problems of attempts sponsored by the International Civil Aviation Organization (ICAO) to develop tests of Aviation English in order to guarantee the qualifications of international pilots to communicate with ground controllers. Once the ICAO had decided on the need for assessment in 2004, four years were allowed to develop suitable tests. In 2008, the deadline was moved to 2011. Alderson and colleagues have surveyed the progress made, and found that many of the assessment instruments being developed are inadequate and that there is little evidence that the tests will meet the hoped-for standards.

These studies confirm that there remains a large gap (virtually an abyss) between policy and assessment: The development of a test to evaluate (or implement) a language policy remains a complex challenge and commonly fails or produces damaging side-effects. I suppose the best analogy remains medicine. Though we have built up a gigantic pharmaceutical industry, a huge number of diseases remain untreated or untreatable, and there is strong evidence of the misuse of many drugs by doctors as well as laypeople. Similarly, the belief that existing testing experts and

https://doi.org/10.1017/S0267190509090011 Published online by Cambridge University Press

businesses can easily produce assessment instruments that will painlessly solve problems of implementing policy goals in a wide variety of areas, whether as complex as controlling an education policy, or as simple (seemingly) as certifying the language proficiency of a future citizen or pilot, appears to be based on wishful thinking rather than hard evidence. Looking at language policy and language assessment together makes eminently clear that there remains a great deal to do.

REFERENCES

Davies, A. (1997). Australian immigrant gatekeeping through English language tests: How important is proficiency? In A. Huhta, V. Kohonon, L. Kurki-Suonio, & S. Luoma (Eds.), *Current developments and alternatives in language assessment : Proceedings of LTRC 96* (pp. 71–84). Jyväskylä, Finland: Kopijyva Oy, University of Jyväskylä.

Edgeworth, F. Y. (1890). The element of chance in competitive examinations. *Journal of the Royal Statistical Society, 53*, 644–663.

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

McNamara, T. (2005). 21st century shibboleth: Language tests, identity and intergroup conflict. *Language Policy, 4*(4), 351–370.

Milani, T. M. (2008). Language testing and citizenship: A language ideological debate in Sweden. *Language and Society, 37*(1), 27–59.

Shohamy, E. (1993). *The power of tests: The impact of language tests on teaching and learning*. Washington, DC: National Foreign Language Center.

Spolsky, B. (1967). Do they know enough English? In David Wigglesworth (Ed.), *ATESL selected conference papers*. Washington, DC: NAFSA Studies and Papers, English Language Series.

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

https://doi.org/10.1017/S0267190509090011 Published online by Cambridge University Press