CHAPTER ONE

Introduction

Language testing has a long history as a social practice but only a relatively short one as a theoretically founded and self-reflexive institutional practice. When language testing became institutionalized in the second half of the 20th century, it did so as an interdisciplinary endeavor between Applied Linguistics and psychometrics. In fact, psychometrics became the substrate discipline, prescribing the rules of measurement, and language was virtually poured into these preexisting psychometric forms. Psychometrics has remained the basic foundation for language testing: it has the appeal of its rigorous methods and anchorage in psychology, a field much older than applied linguistics, with a correspondingly wider research base. However, through marrying itself to psychometrics, language testing has obscured, perhaps deliberately, its social dimension. This is not to say that psychometrics has no awareness of the social context of assessment, but mainstream work in the area is firmly oriented toward assessment instruments with little systematic research on broader issues. In fact, the most socially motivated research in psychometrics on bias and differential item functioning still generally follows an ideal of improving tests so that they measure more precisely and more purely the construct under investigation.

Within the practice of testing, it is, of course, necessary to acknowledge the contribution that psychometrics, despite its limitations, has made and continues to make to socially responsible language testing. Given that modern societies will necessarily use

tests, adherence to the fundamental rules of what makes a "good test" is indispensable for ensuring the quality of assessments and the defensibility of conclusions drawn from them. Following the rigorous prescriptions of psychometrics and measurement theory inherently serves to reduce bias and unwanted social impact. However, and this is a core tenet of our argument, a psychometrically good test is not necessarily a socially good test. In fact, there is no correlation between the two: Any assessment can have far-reaching and unanticipated social consequences. It is precisely these social consequences and the general lack of systematic research on them that we are concerned with in the present volume. This is not a new concern; in fact, the advent of modern testing in the 19th century was accompanied by a vigorous critique of its social and educational implications, as eloquently pointed out by Spolsky (1995) in relation to language testing, reflecting a theme recognized within the more general educational literature represented by, for example, Madaus and Kellaghan (1992). This critique became subdued, certainly within language testing, following the triumph of psychometrics in the 1950s, and this volume represents in a sense only the revoicing of an old issue.

The existence of a social dimension of assessment might well be more striking in language testing than in assessments measuring general cognitive abilities because language is a social medium and any measurement of it outside of a social context tends to be at odds with the increasing acceptance of social models of language within Applied Linguistics more generally. Of course, language tests for a long time ignored the social use dimension of language and followed traditional psychometric methods of measuring isolated pieces of grammar and vocabulary knowledge. However, with the rise of the communicative competence paradigm in second language teaching, the measurement of the ability to use language in social contexts has become increasingly important, as indicated by recent work on oral proficiency interviews (Brown, 2005; Young & He, 1998), performance tests (McNamara, 1996), and tests of pragmatics (Liu, 2006; Roever, 2005). However, here, too, the individualist, cognitive bias in

psychometrics has obscured or even hindered the issues at stake in these assessments, as our discussion shows.

Testing social aspects of language use is only one facet of the social dimension of language tests. The much broader consideration is the role and effect of language testing in society. One of the most pervasive functions and consequences of language tests throughout history is their use as sorting and gatekeeping instruments. These tests, imposed by a more powerful group on a less powerful one, tend to have severe, life-changing, and not infrequently life-ending consequences for their takers, from the Biblical shibboleth test followed by the immediate slaughter of those failing it to the modern language tests that form a precondition for promotion, employment, immigration, citizenship, or asylum. Although the consequences of failure might be less immediate and deadly for the modern tests (except, perhaps, for refugees), they are, nonetheless, high-stakes situations for test takers with the power to define their identity: Passing the test means that one can become a civil servant or a Dutch citizen, incorporating a new facet in one's identity. Also, although modern language tests are frequently administered not by sword-wielding guards but in an orderly, regulated, and lawful fashion in well-lit, carpeted test centers, they are just as much the result of political processes and value decisions as the shibboleth test was. Who should be allowed to immigrate? What standard of language proficiency should be required to grant someone citizenship or allow them to practice medicine in another country? What deeper social and political values are at stake in the implementation of policy through tests in these areas? Social values and attitudes fundamentally determine the use of tests.

Another area in which the consequences of language testing are particularly and sometimes painfully obvious is educational systems. Government-imposed standards and test instruments resulting from them strongly impact what is taught and have a reverberating effect throughout the system, an issue recognized in response to the introduction of such practices within state-sponsored education in the 19th century, itself reflecting an older

tradition of concern dating to the 15th century at least (Madaus & Kellaghan, 1992). From the Common European Framework, which controls language testing and teaching throughout Europe, to the No Child Left Behind Act in the United States, with its special emphasis on English as a second language populations and its grave consequences for various stakeholders, testing in educational systems continues to raise issues of the consequential and social effects of testing, most of which remain the subject of only scant contemporary reflection.

1.1 Structure of This Volume

Our argument is framed in relation to contemporary approaches to validity theory, which we sketch in chapter 2. We describe Messick's (1989) view of validity and validation, including his conceptualization of test consequences, which establishes the social effects of test use as a part of the overall argument for or against a test's construct validity. We consider Messick's influence on his successors: Mislevy's influential Evidence-Centered Design framework (Mislevy, Steinberg, & Almond, 2002, 2003), which implements Messick's approach but focuses on test construction and ignores the consequential facet of validity, in contrast with the work of Kane (Kane, 1992, 2001; Kane, Crooks, & Cohen, 1999), which places decisions based on test scores centrally in his discussion of test validity. We then discuss validity theory in language testing, especially the work of Bachman (Bachman, 1990, 2004; Bachman & Palmer, 1996), in which the influence of Messick and, more recently, Kane is very evident. We argue that despite the creative and sophisticated nature of such discussions, the social role and function of tests is inadequately conceptualized within them.

In chapter 3 we explore the testing of social aspects of language competence. Our argument here is that the individualist and cognitive bias of psychometrics and the linguistics that most neatly fitted it has meant that 40 years after the advent of the communicative movement, the field still encounters difficulties in conceptualizing and operationalizing the measurement of the

social dimension of language use, a problem that was recognized more or less from the outset of the testing of practical communicative skill. We support our argument by considering tests of second language (L2) pragmatics and oral proficiency interviewing. These two measures contrast in interesting ways: Pragmatics tests are rarely used, but their theoretical basis is the subject of much discussion, whereas oral proficiency measures are frequently used with comparatively little reflection. Only four larger measures of L2 pragmatics exist, all focused on English as a target language. These instruments measure speech acts for Japanese learners of English (Hudson, Detmer, & Brown, 1995) or Chinese learners of English (Liu, 2006), as well as implicature (Bouton, 1994), and the only test battery in this area covers routines, implicatures, and speech acts (Roever, 2005). Tests of pragmatics struggle with the necessity of establishing adequate context while keeping the test practical in administration and scoring. Also, the traditional instrument of the discourse completion test (DCT) has been shown to be problematic and there are serious concerns about tearing speech acts out of the context of coherent discourse and atomizing them for testing purposes. Oral Proficiency Interviews take place in the domain of face-to-face interaction, which has been the subject of intense research and theorizing, particularly in the study of language in interaction or Conversation Analysis. One of the central findings of this research is that performance is inherently social and co-constructed, not simply a projection of individual competence; this means that there can be no reading back of individual competence from the data of performance. Schegloff (1995, p. 192) called for a recognition of "the inclusion of the hearer in what [are] purported to be the speaker's processes." The implications of this position for a project that aims at the measurement of what are understood to be individualized cognitive attributes are still being recognized and worked through. Studies are now emerging exploring this issue empirically in detail—for example the research of Brown (2003, 2005), which shows the implication of interlocutor and rater in the scores attributed to individual candidates, a point that had

been raised in principle many years before, for example, in comments by Lado at a symposium in Washington in 1974 (Jones & Spolsky, 1975).

In chapter 4 we discuss ways in which traditional psychometric approaches to testing have handled certain aspects of the social context of assessment. Within this tradition, the issue is conceptualized as one of social fairness, and attention is primarily given to avoiding bias in favor of or against test takers from certain social groups, particularly minorities. Investigation of the psychometric properties of items is done in order to detect items that function differentially for different groups. A variety of statistical techniques have been employed to detect differential item functioning (DIF), ranging from nonparametric cross-tabulations to complex model comparisons based on item response theory. We will describe some of these approaches, together with generalizability theory and multifaceted Rasch measurement, both of which can detect unwanted influences of test-taker background variables at the level of the whole test. Although all of these techniques help detect bias in items or tests, they do not offer much insight into the causes of such unwanted item or test functioning. We discuss this as a shortcoming of traditional DIF and bias analyses and we critically review some of the value decisions made in setting up these analyses.

Although bias avoidance recognizes a social dimension in assessment, it is still fundamentally constructed as a psychometric enterprise. In chapter 5 we consider other aspects of conventional approaches to dealing with the social embeddedness of testing, in the form of procedures of fairness review and the promotion of codes of ethics. Fairness review is a formal process that some test makers operate with the purpose of identifying and eliminating possibly biased items during the test construction process. However, as we show, fairness review is also a response to social and societal pressures on test makers and serves a parallel function of keeping test content uncontroversial to ensure wide acceptance of scores. Similarly, codes of ethics serve as a way for the testing profession to provide transparency to stakeholders and reassure

them of the profession's commitment to ethical conduct. Although there are no effective sanctions for breaches of the ethics code, such codes are useful for guiding ethical decisions and protecting testers from stakeholder pressures to take actions that contravene professional conduct.

In chapter 6 we broaden our scope and discuss ways in which language tests are involved in the construction and defense of particular social identities. Because language indexes social group membership, language tests can be used to classify test takers and this classification is frequently motivated within potentially or actually conflicted intergroup settings and is thus highly consequential. We describe various examples of this social practice from ancient history to the present day, including the Biblical shibboleth test, the Battle of Worcester between Cromwell and Charles II in 1651, the massacre of Haitians in the Dominican Republic in 1937, and the Canadian Mounted Police's "fruit machine" to identify (and exclude) homosexual recruits in the 1950s. We discuss specifically the modern role of language tests in citizenship and immigration contexts, in which they are used to adjudicate the claims of asylum seekers, control legal immigration, and grant or refuse full membership in a society as a citizen. We use this discussion to reflect on the use of language tests as instruments of subjectivity in modern societies.

We then turn in chapter 7 to the pervasive use of tests in educational settings. Standards have long been used to enforce social values throughout educational systems, and tests are the instruments that implement and operationalize such standards. We provide examples of the impact of values and standards in educational systems through the rejection by teachers of an oral component in senior high-school entrance examinations in Japan, as well as in a very different setting, the impact of the Common European Framework of Reference for Languages on language teaching throughout Europe. Increasingly, tests are also used as accountability measures with potentially drastic consequences for various stakeholders and entire educational systems, a use that was first recognized as early as the 15th century in

Europe. This is the case for the No Child Left Behind Act (NCLB) in the United States, where test results determine the fate of entire schools. The test-driven nature of the NCLB has unforeseen collateral effects on various areas of school education in the United States, and the debate about this resurrects debates that have been going on since at least the 19th century.

Finally, in chapter 9 we discuss the implications of our discussion for research and training, most notably a need to devote more research to the social dimension of language testing and to ensure that academic training in language testing balances an appreciation of the achievements of the psychometric tradition and a greater awareness of the social dimensions of language testing.

1.2 Concluding Statement

We feel that language testing is beyond the teething stage and ripe for a broader view of assessment and its social aspects. This includes coming to grips with the theoretical and practical difficulties involved in testing the social side of language use, but testers need again to engage in debate on the increasingly consequential application of their tests and to reflect on test use both during test development and when tests are operational. Language testing has a real impact on real people's lives, and we cannot cease our theoretical analyses at the point where the test score is computed. Just like language use, language testing is and has always been a social practice; the very power of tests has a mesmerizing effect on consciousness of their social character. This volume aims to reawaken that consciousness.