Annual Review of Applied Linguistics (2000) **20**, 147–161. Printed in the USA.

ASSESSMENT AND TESTING

Caroline Clapham

INTRODUCTION

In this brief article, I discuss the relationship between language testing and the other sub-disciplines of applied linguistics and also the relationship, as I see it, between testing and assessment. The article starts with a brief exploration of the term 'applied linguistics' and then goes on to discuss the role of language testing within this discipline, the relationship between testing and teaching, and the relationship between testing and assessment. The second part of the article mentions some areas of current concern to testers and discusses in more detail recent advances in the areas of performance testing, alternative assessment, and computer assessment. One of my aims in this article is to argue that the skills involved in language testing are necessary not only for those constructing all kinds of language proficiency assessments, but also for those other applied linguists who use tests or other elicitation techniques to help them gather language data for research.

APPLIED LINGUISTICS AND THE ROLE OF ASSESSMENT

It is usually the case with new disciplines that they go through periods of adjustment as the limits of the discipline are realigned. Applied linguistics is going through such a stage at present as its scope widens and its subfields start to impinge on those of other disciplines.

The term 'applied linguistics' appears first to have been used in the late 1940s when the discipline embraced the teaching and learning of second and foreign languages (Johnson and Johnson 1998), but since then the discipline has expanded to cover a wider range of sub-disciplines or 'subfields' as they are called by Bachman and Cohen (1998). In 1980, Henry Widdowson said, "…applied linguistics yields descriptions which are projections of actual language which explore linguistic theory as illumination…" (p. 169), and in 1997, Chris Brumfit

defined applied linguistics as "…the theoretical and empirical investigation of real-world problems in which language is the central issue" (p. 93). (See also Ben Rampton's [1997] introduction to the special issue of *Applied Linguistics* in which Brumfit's article appeared; this issue of *Applied Linguistics* focused on the concept of 'applied linguistics.')

In 1999, Richard Hudson, in an e-mail to the LAGB (Linguistics Association of Great Britain) listserve, said:

> The main distinguishing characteristic of AL (Applied Linguistics) is its concern with professional activities whose aim is to solve 'real-world' language-based problems, which means that research touches on a particularly wide range of issues—psychological, pedagogical, social, political and economic as well as linguistic. As a consequence, AL research tends to be interdisciplinary (Hudson 1999).

Language assessment plays a pivotal role in applied linguistics, operationalizing its theories and supplying its researchers with data for their analysis of language knowledge or use. It has itself become a sub-discipline of applied linguistics, which is in some ways unfortunate, since it is has tended to become compartmentalized (Bachman and Cohen 1998) and does not interact as much as it should with the other sub-disciplines. Bachman and Cohen (1998) decry the compartmentalization of second language acquisition (SLA) and language testing, saying that most mainstream researchers in the two sub-disciplines are unaware of research taking place in the other. However, they hope that this lack of awareness may now be changing. (For other discussions of the relationship between testing and SLA, see Shohamy 1998, Upshur and Turner 1998.)

A key point, which perhaps not all applied linguists appreciate, is that language testing is by no means limited to assessing the linguistic proficiency of L2 students. Many areas of linguistic research use elicitation instruments to gather data, and these instruments often take the form of tests or tasks (see, for example, Robinson 1997, Skehan and Foster 1999). If the results of such data elicitation techniques are to be credible, they need to be prepared with as much rigor as proficiency tests, and they therefore have to be valid and reliable (Alderson and Banerjee in press). Crudely, a valid elicitation technique is one that accurately elicits what it is intended to elicit, and a reliable technique is one that produces consistent results. (For discussions of validity, see Chapelle 1999, Messick 1989; 1996, Shepard 1993; and for comments on the relationship between validity and reliability, see Moss 1994.) Although Messick (1989) subsumes reliability under validity, since any valid measure must, by definition, be reliable, it is useful here to distinguish between validity and reliability since the assessment of an instru-ment's reliability is often neglected in elicitation procedures. If research is to have credibility, data gathering instruments must not only be carefully designed to ensure that they will elicit the type of language required, they must also be pre-tested to check that the measures do indeed elicit such language practice and that any rating

or coding system is workable and capable of producing consistent results (see North and Schneider 1998). Since it is generally expected that subjects in an investigation produce similar kinds of language regardless of when the task is done, and that this language can be analyzed in a similar manner regardless of when and by whom it is assessed or coded, the reliability of the elicitation techniques must be given careful consideration.

ASSESSMENT AND LANGUAGE TEACHING

There has been much discussion about how language testing fits into applied linguistics and how it relates to language teaching (see, for example, Bachman and Palmer 1996). In general, it seems clear that "…language testing benefits from insights from applied linguistics as a discipline…" (Alderson and Clapham 1992:164) but that it is sometimes necessary for testing to lead the way:

> We believe that language testers can serve linguistic theory by examining the way in which their tests work, how their different components interrelate, and what they reveal about candidates' language proficiency. Insights from such an analysis of test results should contribute to the development of a better understanding of what is involved in knowing and using language (1992:164).

It seems, indeed, that each affects the other: Methods of assessment may affect teaching in the classroom (Cheng 1997, Wall 1996; 1997), while new theories of language learning and teaching lead to changes in testing practices (Spolsky 1995).

THEORIES OF LANGUAGE TESTING

With the advent of communicative teaching in the late 1970s, there was a need for testers to devise new theories of language testing. Canale and Swain (1980), whose model applied to both teaching and testing second and foreign languages, included grammatical, sociolinguistic, and strategic competence in their description of the domains of language use. In 1990, Bachman added psychophysiological mechanisms and proposed four components in his model: grammatical, textual, illocutionary, and sociolinguistic competence. Bachman and Palmer (1996) elaborated on this model further to include both affective and metacognitive factors. Bachman and Palmer's model of communicative language ability is used as the theoretical basis for tests such as the International English Language Testing System (IELTS) test, and it provides the basis for many current research projects (e.g., Hasselgren 1998). (See McNamara 1996 for a discussion of language testing models.)

ASSESSMENT AND TESTING

The term 'assessment' is used both as a general umbrella term to cover all methods of testing and assessment, and as a term to distinguish 'alternative assessment' from 'testing.' Some applied linguists use the term 'testing' to apply to the construction and administration of formal or standardized tests such as the Test of English as a Foreign Language (TOEFL) and 'assessment' to refer to more informal methods such as those listed below under the heading 'alternative assessment.' For example, Valette (1994) says that 'tests' are large-scale proficiency tests and that 'assessments' are school-based tests. Intriguingly, some testers are now using the term 'assessment' where they might in the past have used the term 'test' (see, for example, Kunnan 1998). There seems, indeed, to have been a shift in many language testers' perceptions so that they, perhaps subconsciously, may be starting to think of testing solely in relation to standardized, large-scale tests. They therefore use the term 'assessment' as the wider, more acceptable term.

Since, for the remainder of this article, I wish to comment on differing attitudes between 'testers' and 'assessors,' I shall use the term 'testers' for those who concern themselves with requirements of validity and reliability, and 'assessors' for those who are not consciously guided by such constraints. I must emphasize, though, that while I am giving the two terms distinct meanings, I do not think that there is a fundamental difference between them, and in other publications (e.g., Clapham 1997; to appear), I use the two terms interchangeably.

Unfortunately, although 'assessors' and 'testers' have the same aims, there is less dialogue than there should be between them, possibly because many of them tend to think of 'testing' and 'assessment' as being categorically different (see Hill and Parry 1994) instead of being on a continuum with at one end those 'testers' who deliver carefully validated multiple choice tests, and at the other end, 'assessors' who prepare real-life tasks for their students and candidates but who do not concern themselves with how well these tasks actually work. 'Assessors' appear to distrust extreme 'testers' because they feel that these 'testers' are so wedded to the numerical analysis of data that they are not sufficiently concerned with the content and administration of their tests. Such 'assessors' tend to be concerned that these 'tests' are not 'communicative,' and that they may lead to negative washback (Brown and Hudson 1998). In contrast, many 'testers' are concerned with the fact that, although the 'assessors'' methods of assessment may be novel and interesting, the tasks are not pre-tested to see whether they work as intended or whether the assessments can be delivered and marked in a consistent manner. In short, they distrust 'assessors' because 'assessors' do not appreciate the importance of investigating the validity and reliability of their instruments. Brown and Hudson (1998) quote Huerta-Macias (1995) who says that it is unnecessary to evaluate the validity and reliability of methods of alternative assessment because they are already built into the assessment process. Brown and Hudson make the point that it is not enough to build validity and reliability into the measures; the

measures must also be trialed to see whether or not they are valid and reliable in practice. (See also Johnstone [in press] and Rea-Dickens [in press].)

Another source of distinction between 'tests' and 'assessments' is that some educators and applied linguists feel that 'high stakes' tests, which have a direct bearing on students' immediate futures, need to have validity and reliability built into them, but that 'low stakes' tests such as classroom tests, which do not have such an obvious impact on students' futures, do not (Davidson, *et al*. 1997). This division of tests into high- and low-stakes types seems to me to be misguided: Tests do not fall neatly into one or the other category. Whether a test is 'high-' or 'low-stakes' is surely a question of degree; a test may be deemed to be more 'high-stakes' than another if students' futures are more clearly at stake, but *all* tests which assess students' proficiency levels, whether in the examination hall or the classroom, are in reality high-stakes. Even if the results of a classroom test do not affect a student's immediate future, the results may become self fulfilling; for example, a student with a low score may be considered by both teacher and student to be a poor language learner, and this may have a damaging effect on the student's future performance. If students are to have an accurate idea of their proficiency, they should, where possible, be given tests which are valid and reliable. Even classroom tests, therefore, should, at least from time to time, be checked to see whether the skills being assessed are those intended, whether the marking scheme is appropriate and can be used consistently, and whether the results tally with other views of the student's proficiency.

This apparent dichotomy between 'testers' and 'assessors' has, however, become less marked as the move towards 'authenticity' (see Bachman 2000) in text and task has led 'testers' towards the increased use of performance tests. In addition, there is a trend at present, perhaps partly due to the influence of Messick (1989), for 'testers' to be more concerned with the construct validity than simply with the reliability of their measures. It is also the case, possibly due to the influence of post modernism, that many 'testers' are rejecting the positivist principle that there is an "…independently existing reality that can be discovered (or measured) using objective, scientific methods..." (Hamp-Lyons and Lynch 1998). The desire to question former 'truths' has led many 'testers' to trust standard statistical procedures less than they used to. Some 'testers' are also now expressing concerns about the ethicality of testing (see Kunnan in press).

CURRENT AREAS OF CONCERN IN TESTING

Areas that are attracting attention in the testing literature at present have, in a number of cases, been the subject of recent *ARAL* reviews. Performance testing was covered by Shohamy (1995), alternative assessment by Hamayan (1995), advances in the use of the computer for testing by Chalhoub-Deville and Deville (1999) and the interface between tasks and assessment by Skehan (1998) and McNamara (1998). Other areas of current interest include test washback (Alderson and Wall 1993; 1996, Wall 1997) and the ethics of language testing (Davies 1997,

Hamp-Lyons 1997, Kunnan in press, Norton 1997). (For more about these and other areas of current concern, see Bachman 2000, Brindley in press; see also Clapham and Corson 1997.) In the remainder of this article, I will update the *ARAL* articles on performance testing, alternative assessment, and the use of computers for testing. I will, at the same time, relate these three areas to real or imaginary differences between 'testers' and 'assessors.'

1. Performance testing

As Shohamy (1995) points out, alternative assessment and performance testing have much in common. Indeed, the major difference between the two seems to be that performance testers agonize about the validity and reliability of their instruments while alternative assessors do not (Hamayan 1995). Both, however, are concerned with asking students to create or produce something, and both focus on eliciting samples of language which are as close to real life as possible (see Kormos 1999, Lynch and McNamara 1998, Papajohn 1999, Upshur and Turner 1998). McNamara (1996) states that a defining characteristic of performance testing is that "the assessment of the actual performances of relevant tasks are required of candidates, rather than the more abstract demonstration of knowledge, often by means of paper-and-pencil tests" (McNamara 1996:6; see also McNamara 1997). The same could well be said of methods of alternative assessment.

2. Alternative assessment

'Alternative assessment' is one of the terms used to refer to informal assessment procedures such as those often used in the classroom. Typical examples of such methods involve portfolios (Hamp-Lyons 1996), learner diaries or journals (Genesee and Upshur 1996), and interviews with teachers (Genesee and Upshur 1996). Such procedures may be more time-consuming and difficult for the teacher to administer than 'paper-and-pencil' tests, but they have many advantages. They produce information that is easy for administrators, teachers, and students to understand; they tend to be integrated; and they can reflect the more holistic teaching methods being used in the classroom (Hamayan 1995). A problem with methods of alternative assessment, however, lies with their validity and reliability: Tasks are often not tried out to see whether they produce the desired linguistic information; marking criteria are not investigated to see whether they 'work'; and raters are often not trained to give consistent marks (see Brown and Hudson 1998). As Hamayan (1995) says, such alternative methods of assessment will not be considered to be part of the mainstream of language assessment until they can be shown to be both valid and reliable.

Methods of alternative assessment are not, of course, solely used for classroom purposes. Many alternative assessment schemes are initiated by local or national government agencies (see Brindley 1998, Stansfield 1994) with aims such as comparing students' levels of linguistic ability or comparing levels of instruction

across institutions. Governments (sometimes in a hurry) wish to use the results for their own aims (Brindley 1998). They do not appreciate the need for careful trialing, and therefore they do not always allow researchers the time to try out the assessment instruments. Some types of assessment used in these circumstances may have high face validity—they may look excellent to the uninformed—but they may be marred by inappropriate marking schemes and rating inconsistencies. Trialing such measures is essential if the final tools are to be valid and reliable, but unfortunately such pre-testing and subsequent editing of materials takes time, and time is often in short supply. Inevitably, the ensuing tasks and marking schemes are not valid and reliable (Brindley 1998), and such schemes, therefore, launched with much fanfare, may produce invalid results which are unfair to students and teachers alike. If all assessors appreciated the importance of trialing in the construction of all assessment procedures, they might be able to work together to tell governments and other funding bodies that test development requires more time if assessors are to devise satisfactory assessment instruments.

The present move towards the inclusion of more performance testing in examinations (see for example the University of Cambridge Local Examinations Syndicate examinations) is also likely to bring 'testing' and 'assessment' closer together, but, unfortunately, because of financial and practical constraints, some large-scale tests are likely to remain much as they are, with uncontextualized, multiple choice items. There is a danger that the authors of such large-scale tests will be marginalized, and that their thorough work on pre-testing their items will not be appreciated because of perceived weaknesses in test content. Indeed, it is possible to envisage a day when performance testers and alternative assessors align themselves against the producers of large-scale tests. Such a division must not be allowed to occur.

3. Advances in computer assessment

It is probably too early to say whether advances in the use of computer technology for testing will have a positive or a negative effect on testing and whether computer administered tests will be distrusted by 'assessors.' Until recently, computer testing tended to fossilize existing objective testing methods because objectively marked items such as multiple-choice questions and gap-filling tasks were straightforward to answer on the computer and were easy to mark mechanically. Any attempts to introduce interesting new methods of assessment and testing were foiled by limitations in the memory size and processing speed of the computers. The move of the TOEFL towards computer based testing too has, at least in the short run, extended the use of multiple choice and other easy-to-mark objective items in computer tests. However, it seems that the promised testing revolution may at last be on its way. The expanding use of video conferencing and video interviews, the comparative ease with which videos and listening extracts can now be downloaded from the Web, the improvement in the computer's ability to recognize sounds and letters, the expanding uses of language corpora for teaching and testing, and the increasing and more rapid storage capacity of modern

computers are all widening the scope of computer administered tests (see Burstein, *et al*. 1996, Drasgow and Olson-Buchanan 1998, Ordinate Corporation 1998). These advances will not only increase the efficiency of standardized tests, but will increase the scope of other elicitation techniques.

One project that has the potential to produce interesting and yet easy-to-deliver-and-mark tests is DIALANG, which aims to produce diagnostic tests in 14 different European foreign languages (DIALANG 1997). The tests will be delivered on the world wide web; they will be computer adaptive (see Chalhoub-Deville and Deville 1999); and students, after taking their chosen test, will receive instant diagnostic information about the strengths and weaknesses of their performance. At present, compositions written for DIALANG will be marked by hand, but this may not always be the case as there are now research projects looking into the computer marking of tests of language production (see Burstein and Chodorow 1999). Although it is hard to imagine computers ever replacing human markers, and we might not wish them to, it is possible that they will take some of the drudgery away from subjective marking and will thus make the marking task easier and more interesting for the raters.

There will thus, in the near future, be great advances in computer testing. However, we cannot yet know whether or not there is still a danger that the use of computers will limit rather than expand the kinds of tests that will be used.

CONCLUSION

One of my aims in this article has been to discuss the need to bring 'testers' and 'assessors' closer together, and I expect that over the next decade any differences between them will become less and less marked. 'Testers,' I hope, will become more open to ideas of different kinds of assessment, and 'assessors' will be more willing to accept that even the most carefully designed task or set of marking criteria needs to be trialed. If 'testers' and 'assessors' can each see that they are aiming for the same goal, they will perhaps start a dialogue which might transform both tests and assessments. Similarly, on a larger scale, the basic issues of testing and assessment are important in all areas of applied linguistics that call on the use of elicitation techniques to collect data. All such elicitation techniques should be reliable and valid. Understanding this connection, however, will only be possible if all putative applied linguists (and all potential language teachers) are introduced during their training to the vital tenets of testing so that they can be more critical of the elicitation techniques they use.

ANNOTATED BIBLIOGRAPHY

Alderson, J. C. and D. Wall. 1993. Does washback exist? *Applied Linguistics*. 14.115–129.

This article opened language testers' eyes to the fact that, although the expression 'washback' was freely used in language teaching and testing communities, and although there was plenty of anecdotal evidence as to its existence, there had been few serious attempts to define it or to investigate its existence and possible effects. The authors reported on the few existing studies of language testing washback and listed fifteen washback hypotheses which, they said, should be investigated. This article led to a flurry of investigations into test washback, and the results of the first of these studies are now being published.

Bachman, L. F. 2000. Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing*. 17.1.

This is a state-of-the-art article on language testing in the year 2000. After a brief overview of testing in the 1980s, Bachman comments on some of the main areas of interest in the 1990s: criterion referenced measurement, generalizability theory, item response theory, structural equation modeling, qualitative research approaches, testing cross-cultural pragmatics, testing language for specific purposes, testing vocabulary, computer based assessment, and research into factors that affect performance. He also discusses recent theories of language testing and the concept of communicative testing. He concludes his overview of the 1990s with discussions of test washback and test ethics, and then turns to what lies ahead.

Bachman, L. F. and A. Cohen. 1998. Language testing—SLA interfaces: An update. In L. F. Bachman and A. Cohen (eds.) *Interfaces between second language acquisition and language testing research*. Cambridge: Cambridge University Press. 1–31.

This chapter serves as a useful introduction to the book, *Interfaces between second language acquisition and language testing research*. It discusses the reasons why SLA and language testing were for some time viewed as totally distinct, and it gives reasons why in recent years the two fields seem to have moved closer together. Bachman and Cohen describe areas of common interest between SLA and language testing and make recommendations for future joint areas of research.

Bachman, L. F. and A. Palmer. 1996. *Language testing in practice*. Oxford: Oxford University Press.

This book, which is a key textbook in language testing at present, is composed of three parts: Part 1 discusses the conceptual basis of test development and includes the Communicative Language Ability model, which is used as the basis of much current research in language testing; Part 2 describes the stages of language test development; and Part 3 describes a set of ten illustrative test development projects ranging from a placement test used in a U.S. university to a syllabus based test for primary school children.

Brindley, G. 1998. Outcomes-based assessment and reporting in language learning programmes: A review of the issues. *Language Testing*. 15.45–85.

In this article, Brindley discusses national standards, frameworks, and benchmarks of various kinds. He shows that the introduction of such systems has sometimes been problematic because of political, practical, and technical factors. Brindley discusses what the problems are and makes suggestions for how future systems might be more successfully implemented.

Brown, J. D. and T. Hudson. 1998. The alternatives in language assessment. *TESOL Quarterly*. 32.653–675.

The main purpose of Brown and Hudson's article is to help language teachers decide what types of language tests they should use in the classroom. The article discusses the concepts of validity and reliability, and lists the different kinds of tests that teachers might use. The authors group tests under various headings such as 'selected response' and 'performance assessments' and describe the advantages and disadvantages of each type.

Clapham, C. and D. Corson (eds.) 1997. *Language testing and assessment, Vol. 7. The encyclopedia of language and education*. Dordrecht, Holland: Kluwer Academic.

This volume contains 29 chapters on different aspects of first and second language testing and assessment. Each chapter presents a state-of-the-art description of one aspect of language assessment and provides a bibliography of about 30 references for future researchers in the field. The book is divided into four sections covering the testing of individual skills, methods of assessment, quantitative and qualitative approaches to test validation, and the ethics and effects of testing and assessment.

McNamara, T. 1996. *Measuring second language performance*. London: Longman.

This highly readable book starts with a discussion of communicative testing and defines what McNamara means by performance testing. This discussion is followed by a valuable discussion of the influential models of communicative language testing that have been devised since Dell Hymes introduced his theory of communicative competence in 1972. In the second part of the book, McNamara starts by describing a performance test, the Occupational English Test, and then devotes most of the rest of the book to a discussion of the use of Rasch multi-faceted measurement for research into the assessment of second language performance.

Messick, S. 1996. Validity and washback in language testing. *Language Testing*. 13.241–256.

This article focuses on washback and the consequential aspects of construct validity. Messick relates washback to his overall conception of validity and, in doing so, explains his ideas about validity more clearly and concisely than he did in his seminal 1989 article.

Shepard, L. 1993. Evaluating test validity. *Review of Research in Education*. 19.405–450.

This article discusses validity in educational measurement as a whole. Shepard starts by describing the evolution of the concept of validity over time and then devotes the second half of the article to a very clear explanation of Messick's theory of validity. She concludes the chapter with comments on the implications of this reconceptualizing of the term for new test standards.

Wall, D. 1996. Introducing new tests into traditional systems: Insights from general education and from innovation theory. *Language Testing*. 13.334–354.

This article describes several key concepts in educational innovation. The author applies these concepts to the teaching of English as a foreign or second language and relates them to a study she carried out into the washback of a new school examination in Sri Lanka. She shows how the belief that assessment and the curriculum would together affect teaching in the classroom turned out to be misplaced, partly because of discrepancies between the curriculum and the examination, and partly because of a lack of teacher training in the new 'communicative' methodology. In her conclusion, she makes suggestions as to how future investigations into washback should be carried out and how innovations in the classroom might be brought about more successfully.

UNANNOTATED BIBLIOGRAPHY

Alderson, J. C. and J. Banerjee. In press. Impact and washback research in language testing. In C. Elder, *et al.* (eds.) *Experimenting with uncertainty: Essays in honor of Alan Davies*. Cambridge: Cambridge University Press.

_____ and C. Clapham. 1992. Applied linguistics and language testing: A case study of the ELTS test. *Applied Linguistics*. 13.149–167.

_____ and D. Wall (eds.) 1996. *Washback in testing*. [Special issue of *Language Testing*. 13.1.]

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

Brindley, G. In press. Assessment. In R. Carter and D. Nunan (eds.) *The Cambridge ELT companion*. Cambridge: Cambridge University Press.

Brumfit, C. 1997. How applied linguistics is the same as any other science. *Applied Linguistics*. 7.86–94.

Burstein, J. and M. Chodorow. 1999. Automated essay scoring for non-native English speakers. Paper presented at the Association of Computational Linguistics and the International Association of Language Learning Technologies. [Available from http://www.ets.org/research/erater.html]

_____ T. Frase, A. Ginther and L. Grant. 1996 Technologies for language assessment. In W. Grabe, *et al.* (eds.) *Annual Review of Applied Linguistics, 16. Language and technology*. New York: Cambridge University Press. 240–260.

Canale, M. and M. Swain. 1980. Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*. 1.1–47.

Chalhoub-Deville, M. (ed.) In press. *Issues in the computer adaptive testing of reading proficiency*. Cambridge: Cambridge University Press.

_____ and C. Deville. 1999. Computer adaptive testing in second language contexts. In W. Grabe, *et al.* (eds.) *Annual Review of Applied Linguistics, 19. Survey of applied linguistics*. New York: Cambridge University Press. 273–299.

Chapelle, C. 1998. Construct definition and validity enquiry in SLA research. In L. F. Bachman and A. Cohen (eds.) *Interfaces between second language acquisition and language testing research*. Cambridge: Cambridge University Press. 32–70.

_____ 1999. Validity in language assessment. In W. Grabe, *et al.* (eds.) *Annual Review of Applied Linguistics, 19. Survey of applied linguistics*. New York: Cambridge University Press. 254–272.

Cheng, L. 1997. How does washback influence teaching? Implications for Hong Kong. *Language Education*. 11.38–54.

Clapham, C. 1997. Review of Hill, C. and Parry, K. (eds.) *From testing to assessment*. 1994. *Language and Education*. 11.222–223.

_____ To appear. Assessment. In M. Byram (ed.) *Encyclopedia of Language Teaching and Learning*. London: Routledge.

Davidson, F., C. E. Turner and A. Huhta. 1997. Language testing standards. In C. Clapham and D. Corson (eds.) *Language testing and assessment, Vol. 7.*

The encyclopedia of language and education. Dordrecht, Holland: Kluwer Academic. 303–311.

Davies, A. (ed.) 1997. *Ethics in language testing*. [Special issue of *Language Testing*. 14.3.]

DIALANG. 1997. 'DIALANG: A new European system for diagnostic language assessment.' *Language Testing Update*. 21.38–39. [http://www.jyu.fi/DIALANG]

Drasgow, F. and J. Olson-Buchanan (eds.) 1998. *Innovations in computerized assessment*. Mahwah, NJ: L. Erlbaum.

Genesee, F. and J. Upshur. 1996. *Classroom-based evaluation in second language education*. Cambridge: Cambridge University Press.

Hamayan, E. 1995. Approaches to alternative assessment. In W. Grabe, *et al.* (eds.) *Annual Review of Applied Linguistics, 15. Survey of applied linguistics*. New York: Cambridge University Press. 212–226.

Hamp-Lyons, L. 1996. Applying ethical standards to portfolio assessment of writing in English as a second language. In M. Milanovic and N. Saville (eds.) *Performance testing, cognition and assessment*. Cambridge: Cambridge University Press. 151–162.

_____ 1997. Ethics in language testing. In C. Clapham and D. Corson (eds.) *Language testing and assessment, Vol. 7. The encyclopedia of language and education*. Dordrecht, Holland: Kluwer Academic. 323–333.

_____ and B. Lynch. 1998. Positivistic versus alternative perspectives on validity within the LTRC. Unpublished manuscript.

Hasselgren, A. 1998. Small words and valid testing. Bergen, Norway: University of Bergen. Unpublished Ph. D. diss.

Hill, C. and K. Parry. 1994. *From testing to assessment*. London: Longman.

Hudson, R. 1999. E-mail message to the Linguistics Association of Great Britain (LAGB) listserve. [See http://www.phon.ucl.ac.uk/home/dick/AL.html]

Huerta-Macias. 1995. Alternative assessment – Responses to commonly asked questions. *TESOL Journal*. 5.8–11.

Johnson, K. and H. Johnson (eds.) 1998. *Encyclopedic dictionary of applied linguistics*. Malden, MA: Blackwell.

Johnstone, R. In Press. Context-sensitive assessment of modern languages in primary and early secondary education: Scotland and the European experience. *Language Testing*. 17.

Kormos, J. 1999. Simulating conversations in oral proficiency assessment: A conversation analysis of role play and non-scripted interviews in language exams. *Language Testing*. 16.163–188.

Kunnan, A. J. 1998. An introduction to structural equation modeling for language assessment research. *Language Testing*. 15.295–332.

_____ (ed.) In press. *Fairness and validation in language assessment*. Cambridge: Cambridge University Press.

Lynch, B. and T. McNamara. 1998. Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*. 15.158–180.

McNamara, T. 1997. Performance testing. In C. Clapham and D. Corson (eds.) *Language testing and assessment, Vol. 7. The encyclopedia of language and education.* Dordrecht, Holland: Kluwer Academic. 131–139.

_____ 1998. Policy and social considerations in language assessment. In W. Grabe, *et al.* (eds.) *Annual Review of Applied Linguistics, 18. Foundations of second language teaching*. 18.304–319.

Messick, S. 1989. Validity. In R. L. Linn (ed.) *Educational measurement*. New York: American Council of Education/Macmillan. 13–103.

Moss, P. 1994. Can there be validity without reliability? *Educational Researcher*. 23.8.5–12.

North, B. and G. Schneider. 1998. Scaling descriptors for language proficiency scales. *Language Testing.* 15.217–262.

Norton, B. 1997. Accountability in language Assessment. In C. Clapham and D. Corson (eds.) *Language testing and assessment, Vol. 7. The encyclopedia of language and education.* Dordrecht, Holland: Kluwer Academic. 313–322.

Ordinate Corporation. 1998. *PhonePass test validation report*. Menlo Park, CA: Ordinate.

Papajohn, D. 1999. The effect of topic variation in performance testing: The case of the chemistry TEACH test for international teaching assistants. *Language Testing*. 16.52–81.

Rampton, B. 1997. Retuning in applied linguistics. *Applied Linguistics*. 7.3–25.

Rea-Dickens, P. In press. Snares and silver bullets: Disentangling the construct of formative assessment. *Language Testing*. 17.

Robinson, P. 1997. Individual differences and fundamental similarities of implicit and explicit adult second language learning. *Language Learning*. 47.45–99.

Shohamy, E. 1995. Performance assessment in language testing. In W. Grabe, *et al.* (eds.) *Annual Review of Applied Linguistics, 15. Survey of applied linguistics*. New York: Cambridge University Press. 188–211.

_____ 1998. How can language testing and SLA benefit from each other? The case of discourse. In L. F. Bachman and A. Cohen. (eds.) *Interfaces between second language acquisition and language testing research*. Cambridge: Cambridge University Press. 156–176.

Skehan, P. 1998. Task-based instruction. In W. Grabe, *et al.* (eds.) *Annual Review of Applied Linguistics, 18. Foundations of second language teaching*. New York: Cambridge University Press. 304–319.

_____ and P. Foster. 1999. The influence of task structure and processing conditions on narrative retellings. *Language Learning*. 49.93–120.

Spolsky, B. 1995. *Measured words*. Oxford: Oxford University Press.

Stansfield, C. 1994. Developments in foreign language testing and instruction: A national perspective. In C. Hancock (ed.) *Teaching, testing and assessment: Making the connection.* Lincolnwood, IL: National Textbook Company. 43–67.

Upshur, J. and C. Turner. 1998. Systematic effects in the rating of second-language speaking ability: Test method and learner discourse. *Language Testing*. 16.82–111.

Valette, R. 1994. Teaching, testing and assessment: Conceptualizing the relationship. In C. Hancock (ed.) *Teaching, testing and assessment: Making the connection*. Lincolnwood, IL: National Textbook Company. 1–42.

Wall, D. 1997. Impact and washback in language. In C. Clapham and D. Corson (eds.) *Language testing and assessment, Vol. 7. The encyclopedia of language and education*. Dordrecht, Holland: Kluwer Academic. 291–302.

Widdowson, H. G. 1980. Models and fictions. *Applied Linguistics*. 1.165–170.