Experimenting with uncertainty

Essays in honour of Alan Davies
STUDIES IN LANGUAGE TESTING...11
Series editor: Michael Milanovic

Also in this series:

An investigation into the comparability of two tests of English as a Foreign Language: The Cambridge–TOEFL comparability study
Lyle F. Bachman, F. Davidson, K. Ryan, I.-C. Choi

Test taker characteristics and test performance: A structural modeling approach
Antony John Kunnan

Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem
Michael Milanovic, Nick Saville

The development of IELTS: A study of the effect of background knowledge on reading comprehension
Caroline Margaret Clapham

Verbal protocol analysis in language testing research: A handbook
Alison Green

A multilingual glossary of language testing terms
prepared by ALTE members

Dictionary of language testing
Alan Davies, Annie Brown, Cathie Elder, Kathryn Hill, Tom Lumley, Tim McNamara

Learner strategy use and performance on language tests
James Enos Purpura

Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida
Antony John Kunnan

Issues in computer-adaptive testing of reading proficiency
Micheline Chalhoub-Deville
Experimenting with uncertainty
Essays in honour of
Alan Davies

Edited by C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara, and K. O'Loughlin
WRITING OF TESTS

(with apologies to Henry Reed ‘Lessons of the War: 1, Naming of Parts’ (1946))

Today we have writing of tests. Yesterday
We had rater training, and tomorrow morning,
We shall have what to do after marking. But today,
Today we have writing of tests. Niphophila
Tremble like dancers on all of the Dandenong Ranges
And today we have writing of tests.

This is the ASLPR bandscales. And these
Are the IELTS bandscales, whose use you will see,
When you are given your specs. And these are the TOEFL bandscales
Which in your case you have not got. The sprinklers
Arc in the gardens their pulsing mysterious signals
Which in our case we have not got.

Alan Davies

(On the occasion of his retirement from the Language Testing Research Centre, Melbourne 1998)
Contents

Series Editor’s note x
Preface xi

Section One
The contribution of Alan Davies 1

1. Alan Davies and British applied linguistics 2
   Christopher Brumfit

2. Ten years of the Language Testing Research Centre 5
   Tim McNamara

Section Two
Construct definition in language testing 11

3. Communicative language testing: The art of the possible 12
   Henry Widdowson

4. Fossilisation or evolution: The case of grammar testing 22
   Pauline Rea-Dickins

5. The assessment of metalinguistic knowledge 33
   Caroline Clapham

Section Three
Language testing for specific purposes and populations 44

6. Three problems in testing language for specific purposes: Authenticity, specificity and inseparability 45
   Dan Douglas

7. Assessing language skills for specific purposes: Describing and analysing the ‘behaviour domain’ 53
   Elaine Tarone

© Cambridge University Press  www.cambridge.org
8. The assessment of language impairment across 61
language backgrounds
Rosemary Baker

9. A process for translating achievement tests 73
Charles Stansfield and Joan Auchter

Section Four
Judgement in language testing 81

10. Revisiting raters and ratings in oral language assessment 82
Daniel Reed and Andrew Cohen

11. Establishing meaningful language test scores for 97
selection and placement
Patrick Griffin

Section Five
The uses and usefulness of language tests 108

12. Designing and developing useful language tests 109
Lyle Bachman

13. The formative and summative uses of language test data: 117
Present concerns and future directions
Cyril Weir

14. Language assessment and professional development 126
Geoff Brindley

Section Six
Language test impact 137

15. The need for impact studies of L2 performance testing 138
and rating: Identifying areas of potential consequences at
all levels of the testing cycle
Carolyn Turner

16. Impact and washback research in language testing 150
J. Charles Alderson and Jayanti Banerjee
Section Seven
Language testing in its policy context 162

17. Prescribed language standards and foreign language classroom practice: Relationships and consequences
   Rosamond Mitchell

18. Rendering ESL accountable: Educational and bureaucratic technologies in the Australian context
   Helen Moore

19. The policy context of English testing for immigrants
   John Read

20. Testimony from testees: The case against current language policies in sub-Saharan Africa
   Eddie Williams

Section Eight
The ethics of language testing 211

21. Cheating language tests can be dangerous
   Bernard Spolsky

22. Ethics, fairness(es), and developments in language testing
   Liz Hamp-Lyons

23. The ethical potential of alternative language assessment
   Brian K. Lynch

Section Nine
Language testing and SLA 240

24. Quantitative evaluation of vocabulary: How it can be done and what it is good for
   Batia Laufer

25. Some thoughts on testing grammar: An SLA perspective
   Rod Ellis
26. Measuring development and ultimate attainment in non-native grammars
Antonella Sorace and Daniel Robertson

Section Ten
Beyond language testing

27. Fossilisation: Moving the concept into empirical longitudinal study
Larry Selinker and ZhaoHong Han

28. The unbearable lightness of being a native speaker
John C. Maher

Section Eleven
The publications of Alan Davies

Notes on the contributors
Author Index
Series Editor’s note

Alan Davies has been centrally involved in Applied Linguistics and Language Testing for more than thirty years. Over that time he has also worked with UCLES on many occasions, most recently as a consultant, advisor and editor-in-chief of the SILT Volume 7, *A Dictionary of Language Testing*. Alan’s contributions to our work in Cambridge have always been of the greatest help and it is with a sense of honour that we publish this volume as a token of our respect.

Michael Milanovic
Cambridge
September 2000
Preface

This volume pays tribute to Alan Davies’ work in the field of language testing which spans a period of more than thirty years. His interest in this area began almost by accident when, on leave in the UK from an English teaching post in Kenya, he was offered the opportunity of working at Birmingham University on the development of the English Proficiency Test Battery (E. P. T. B.), later known as the ‘Davies’ test, commissioned by the British Council for the selection of overseas students for admission to higher education courses in the UK. His involvement in this project provided the data for his doctoral dissertation ‘Proficiency in English as a Second Language’ which was submitted in 1965. The project gave him the chance to acquire knowledge and skills in educational measurement, but its chief appeal was that it offered him a way to understand and build theory around the problems he had faced when applying English as a mother tongue assumptions to his practice as a teacher of English to L2 learners in East Africa.

It is this interest in language testing not as an end in itself but as a means of exploring or operationalising important issues in applied linguistics that characterises much of Davies’ contribution to the field, together with his constant striving for the right balance between speculation and empiricism (Davies 1992a). He writes in the first issue of the journal *Language Testing* which appeared in June 1984:

*The process of concurrent and predictive validities, the internal analyses and the external comparisons are time consuming but routine. …in the end no empirical study can improve a test’s validity. That is a matter for the construct and content validities. What is most important is the preliminary thinking and the preliminary analysis as to the nature of the language learning we aim to capture…*  
(p. 68)

Davies (1968b) has warned against allowing the necessary preoccupation with the psychometric properties of language tests to override the central issues of language, learning and evaluation which underlie the language testing enterprise. What sets his own books on testing, *Testing and Experimental Methods, Edinburgh Course in Applied Linguistics, Vol 4* (Allen and Davies 1977) and *Principles of Language Testing* (Davies 1990a), apart
Preface

from others in the field is their attempt to bring language testing squarely under the umbrella of Applied Linguistics.

It is important to acknowledge (as do a number of authors in this volume) that Davies’ interests and scholarly endeavours are by no means confined to language testing. He has written extensively about language teaching and learning, sociolinguistics, language planning, ideology and, more recently, about the nature and scope of the Applied Linguistic endeavour (see The publications of Alan Davies at the end of this volume). His originality lies in his capacity to draw these various strands together in his language testing work, to ask unexpected, but always pertinent, questions, to query whether a new trend is necessarily better than, or essentially different from, what has preceded it, and to draw attention to the wider implications of a particular test use or test outcome. We are acutely aware that the focus of this volume has meant that many who would have liked to mark their respect for Davies’ contribution to the broader field of applied linguistics have felt unable to do so in this context. We are nevertheless happy that we have been able to attract contributors (e.g. Widdowson, Section Two; Tarone, Section Three; Sorace and Robertson, Section nine; Maher, Section ten) who do not see themselves as language testers but who nevertheless recognise the potential or actual links between this field and their own research interests and/or who are prepared to give us the benefit of the outsider’s view.

The papers in this Festschrift have been collected by colleagues at the Language Testing Research Centre where Alan worked for several years before retiring in March 1998. (He remains attached to the Centre as Principal Fellow but has now returned to the University of Edinburgh where he is Honorary Fellow and Emeritus Professor of Applied Linguistics.) Its scope is deliberately broad. It presents a number of different ‘angles’ on language testing, most of which have been touched upon in Alan’s writings over the years. There are 28 papers in all, divided into 10 sections covering issues which range from construct definition in language testing to the design and applications of language tests (including their importance as a means of exploring larger issues in Applied Linguistics) and the consequences (pedagogical, social and ethical) of their use. The papers have been grouped thematically but these groupings should be seen as ad hoc since it was clear from our editorial discussions that a number of alternative categorisations would have been possible.

Section One The contribution of Alan Davies contains two papers, Alan Davies and British applied linguistics and Ten years of the Language Testing Research Centre, which pay direct tribute to Alan’s work over the last three decades in both Britain and Australia. The first, by Christopher Brumfit, outlines Alan’s outstanding record in British applied linguistics and points to the profound ‘humanising influence’ of his work across the whole discipline.
The second, written by Tim McNamara in collaboration with the other editors of this volume, tells the story of a major research initiative in Australia with which Alan was associated and which would not have been possible without Alan’s benign leadership, his prolific scholarly output and his capacity to encourage and inspire his younger colleagues.

The second section, Construct definition in language testing, deals with the controversial (and to Davies absolutely crucial) issue of constructs of language ability (e.g. Canale and Swain 1980, Bachman 1990) and how these are operationalised in language tests. It opens with a paper by Henry Widdowson, Communicative language testing: The art of the possible, which pays homage (through a personal anecdote) to Alan’s admirable and sometimes unnerving capacity to ‘disturb the settled certainties of conventional belief’. Widdowson, in similar vein, proceeds to mount his own challenge to current orthodoxy in language testing. He questions the adequacy of recent models of communicative competence (because of their failure to capture the interrelationships between the various components of competence that they identify) and casts doubt on the possibility that such competence, as currently defined, can be tested at all. Widdowson’s doubt about the feasibility of communicative language testing echoes Davies’ early scepticism in the face of what he saw as over-eager assumptions about the value of CLT.

Naturalism is a vulgar error: all education needs some measure of idealisation, and the search for authenticity in language teaching is chimerical … Testing (like teaching) the communicative skills is a way of making sure that there are tests of context as well as of grammar; testing (and teaching) the communicative skills is not doing something parallel to or different from testing (and teaching) the linguistic skills – what it does is to make sure that they are complete.

(Davies 1978a)

His caution has proved to be well-founded. So too has his prediction that grammar, more powerful in terms of generalisability than any other language feature, would remain central to language testing (ibid). The practice of grammar testing, as Pauline Rea-Dickins reminds us in the second paper in this section, Fossilisation or evolution: The case of grammar testing, is still widespread. Her paper examines the prevalence of tests explicitly focused on grammar in a range of situations, concluding that grammar seems to hold a more prominent position when students are entering ESL/EAP courses than when they exit from them. She suggests that grammar has come to be defined much more broadly than as sentence level accuracy, and this is reflected in the range of integrative methods used to test it – here there are signs of evolution. On the other hand, the reasoning and practices associated with grammar
testing in pedagogic contexts are much less clear, and suggest a more fossilised situation. She urges language testers to provide guidance to the language teaching profession by coming up with better definitions of grammar and by clearly operationalising their constructs.

The next paper, *The assessment of metalinguistic knowledge*, by Caroline Clapham, shows how exacting the task of construct definition can be. Her concern is with the metalinguistic knowledge of undergraduate foreign language learners and with how it can be tested – an important issue given a) the emphasis which university teachers place on explicit grammar teaching and on their students’ understanding of grammatical terms, and b) current controversy in SLA about the role of knowledge about language in acquiring a second language. Clapham’s study, which has since been partially replicated by Davies and colleagues in Melbourne (Elder *et al.* 1997), demonstrates how difficult it is to decide at what point a learner’s understanding can be regarded as adequate. Her results indicate that language knowledge varies according to the task or context in which it is elicited and that what learners appear to understand in one language is not necessarily carried over into another.

In the third section, *Language testing for specific purposes and populations*, we have grouped four papers which deal, in one way or another, with test design issues or, more precisely, with the matching of test tasks with test takers’ particular backgrounds, abilities or needs. The first two are concerned with testing language for specific purposes and take up issues which have been foreshadowed by Davies in his work on the ELTS validation study (Criper and Davies 1988), on a medicine-specific listening test for the Professional and Linguistic Assessment Board (PLAB) (Davies 1986) and on LSP and performance testing generally (Davies and Brown 1990; Davies 1990a, 1995a).

The first paper, by Dan Douglas, *Three problems in testing language for specific purposes: Authenticity, specificity and inseparability*, offers an overview of the thorny problems surrounding LSP test development. Those which Douglas identifies as critical are the following: a) how to achieve the best fit between the test and the relevant domain of inference (= authenticity), b) how to achieve a balance between generalisability of test results and the tailoring of test tasks required to render them appropriate for a particular group of test takers (= specificity), and c) how (if at all) we can identify and measure the relative contributions of language ability and content knowledge to LSP test performance (= inseparability).

Elaine Tarone’s article, *Assessing language skills for specific purposes: Describing and analysing the ‘behaviour domain’*, focuses more specifically on the methodology of specific-purpose testing and proposes a principled approach to the needs analysis stage of test development which draws on analytic frameworks and ethnographic techniques adopted for genre analysis.
These techniques, she suggests, can yield more accurate and meaningful descriptions of language behaviour in context and as such are a necessary first step in the design of both LSP tests and classroom activities. Whether the items or tasks thus derived would necessarily be performance based (in keeping with the central LSP testing position) or more indirect measures of the relevant skills and abilities (as Davies [1986] would have it) remains uncertain.

Another angle on test specificity is the question of how validly a test designed for a specific purpose can measure the performance of a particular subpopulation of test takers – an issue addressed by Davies (1991b) in relation to the performance of children from immigrant backgrounds on basic numeracy tests administered through the medium of English. A similar issue is taken up by Rosemary Baker in her paper, *The assessment of language impairment across language backgrounds*. Baker finds that language tasks of the types commonly used for the assessment of language impairment in age-related disease, when administered in English, can disadvantage patients from non-English-speaking backgrounds and result in diagnostic errors or virtual denial of access to speech pathology services. But Baker also warns against the assumption that non-native speakers of English necessarily perform better in their first languages and identifies a number of problems with direct translations of test content into other languages.

Stansfield and Auchter also deal with the issue of test equivalence across languages in their paper, *A process for translating achievement tests*. They describe the painstaking procedures adopted to translate tests from one language to another in such a way as to ensure that all items are valid and that the resultant instruments measure comparable constructs. Although the authors agree with Baker (above) that translation is not *ipso facto* a guarantee of item equivalence across different languages and cultures, they suggest that strict adherence to pre-established translation guidelines can reduce the risk of construct-irrelevant differences in test performance, and resultant biases in score interpretation.

The fourth section, *Judgement in language testing*, deals with an issue which Davies regards as endemic to all attempts to study and measure language learning, namely, the inevitable subjectivity involved in making decisions about how a test is to be designed, how performances are to be rated and what inferences are to be drawn from test scores (see for example Davies *et al.* 1996b). The section contains two papers, a review article by Daniel Reed and Andrew Cohen on the issue of rater behaviour in language testing and a methodology paper by Patrick Griffin dealing with the notoriously problematic issue of setting cut-offs in language tests (see also Davies 1990b).

The Reed and Cohen paper contains a survey of the relevant literature on (a) the validity of different kinds of rating scales and procedures, (b) the
characteristics of raters, and (c) rater behaviour and its amenability to training. The conclusion reached by the authors is that some of the most crucial questions about enhancing the validity of language test ratings have yet to be answered. They propose a research agenda into rater behaviour which encompasses both quantitative and qualitative methodologies, and advocate the construction of new types of rating scales which reflect test constructs more explicitly.

Griffin’s paper, *Establishing meaningful language test scores for selection and placement*, proposes a practical means of accommodating rater variability and minimising the inevitable uncertainty involved in the setting of cut-scores on language tests through the combined application of the Angoff method and the Rasch partial credit model of measurement. The author however emphasises that the validity of this process rests ultimately on the validity of rater judgements. Griffin argues that meaningful cut-off scores can only be ensured if there is a close fit between the scoring method and the variable being measured, and if the raters concerned are both specialists in the relevant domain and trained in using the scoring procedure.

Section Five, *The uses and usefulness of language tests*, contains three papers centred around the issue of how tests are used and how they can be rendered useful for their intended purposes.

Lyle Bachman, in his paper *Designing and developing useful language tests*, presents us with a theoretically-grounded framework in which considerations of test usefulness permeate the entire testing cycle, from inception to ultimate use. He sees a key element of test authenticity, and hence the construct validity of score interpretations, as being a demonstrable correspondence between the characteristics of target language use tasks and those of the test tasks, a view which appears on the surface to be quite contrary to that held by Widdowson. Further analysis of the two positions may however reveal considerable affinity between Widdowson’s notion of ‘valency’ (see Section Two) and Bachman and Palmer’s definition of ‘interactiveness’ as the extent and type of involvement of the test taker’s individual characteristics in accomplishing a test task (Bachman and Palmer 1996: 25). Both, we suspect, would in the end agree with Davies’ (1990a) conclusion:

*The fundamental argument/debate in language testing over the last 24 years has been basically about the meaning/realization of language behaviour, how best to get at it. The issue is sometimes presented as if there were disagreement about language use. There is not. The disagreement is about the best way to capture control of that language use …*

(1990a: 137)

Preface
The topic of Cyril Weir’s paper, *The formative and summative uses of language test data: Present concerns and future directions*, is highly appropriate to this *Festschrift*, given the number of evaluation projects which Davies has been involved in during the course of his career (e.g. Beretta and Davies 1985, 1986; Davies 1987, 1990a, 1991c). Weir, like Davies, is somewhat circumspect about the role of testing in the evaluation process although for different reasons. Davies makes modest claims about the value of tests in programme evaluation seeing them merely as ‘a way of focusing attention, discussion and planning on the original and on the existing purposes’ (1990a: 116). Weir is more concerned with the qualitative information about learning processes which may be lost or ignored as a result of undue faith in the testing product. Paradoxically, he points out, the reverse is the case in much formative evaluation where in many cases testing is not as central as it should or could be to the classroom monitoring of student language development.

The third paper in this section, *Language assessment and professional development*, by Geoff Brindley, is about test users, rather than test use. It takes on the important task of delineating what these users need to know about language testing (a task which has also been tackled by Davies and his colleagues through their video series: *Mark My Words: Assessing second and foreign language skills* (Davies et al. 1996a) and through the creation of a *Dictionary of Language Testing* (Davies et al. 1999).) Brindley outlines various ways in which social, economic and political forces have influenced policies and practices in educational assessment in recent years and considers the implications for teachers of this changing assessment environment. He concludes by drawing up an agenda for professional development programmes which will enable teachers to perform learner assessments in a competent manner and to be aware of the long-term impact of their testing practices (see Turner, and Alderson and Banerjee, below).

Test effects or test consequences are now widely believed to be a key aspect of test validity and of a tester’s responsibility (Messick 1989) and it is perhaps for this reason that the papers in Section Six, *Language test impact*, demand that greater attention be paid to this phenomenon. Davies, however, while stressing the importance of professional accountability, has warned language testers against defining their responsibilities so broadly that they become unmanageable (1997b: 335–6). Carolyn Turner, in her paper, *The need for impact studies of L2 performance testing and rating: Identifying areas of potential consequences at all levels of the testing cycle*, argues that in order to better define where the profession’s responsibilities start and finish, and also to maximise the beneficial impact of language tests, we should strengthen our understanding of the washback phenomenon through further empirical research.
Preface

J. Charles Alderson and Jayanti Banerjee, while acknowledging the importance of empirical work on test impact, stress the need for greater methodological rigour in its implementation. Their paper, *Impact and washback research in language testing*, concentrates on the validation of instruments used for data gathering purposes. The impact studies reviewed by the authors lacked any adequate treatment of validation issues – hence their recommendation that concepts from the field of language testing be used to develop a conceptual framework within which validity and reliability issues can be investigated. These proposals are illustrated with reference to an ongoing International English Language Testing Service (IELTS) impact study in which the authors are engaged.

In Section Seven, *Language testing in its policy context*, we have grouped papers which place testing or assessment issues in the context of national language policies in five different countries: the United Kingdom, Australia, New Zealand, Malawi and Zambia. The demand by governments for greater accountability in education is a near-global phenomenon (see Brindley, Section Five) and Rosamond Mitchell, in a paper entitled *Prescribed language standards and foreign language classroom practice: Relationships and consequences*, discusses the introduction of ‘language standards’ to foreign language (FL) teaching in England and Wales. The author examines the model of FL development underlying the National Curriculum and suggests that its failure to incorporate ‘growth’-oriented beliefs about language use and interlanguage development is likely to stifle creativity and experimentation by both teachers and learners, thereby seriously hampering the achievement of communicative proficiency outcomes in FL classrooms.

Helen Moore’s paper, *Rendering ESL accountable: Educational and bureaucratic technologies in the Australian context*, is also concerned with the issue of national frameworks for curriculum and assessment, this time in the context of adult migrant education. While conceding (similarly to Mitchell, above) that such state-mandated frameworks are reductionist and tend to induce conformity rather than foster creativity, she argues that they can at the same time offer important educational trade-offs. She demonstrates that the recently introduced Certificates in Spoken and Written English for adult learners of English as a second language, notwithstanding their limitations, provide a powerful authorising mechanism whereby teachers’ professional aspirations can be realised and institutional claims for resources can be legitimised.

John Read’s paper, *The policy context of English testing for immigrants*, documents the social impact of recent policy changes for immigrants to New Zealand including the introduction of a controversial pre-entry English language requirement and a financial penalty for those failing to meet this requirement within a given period of time. His paper is a welcome addition to
recent writings on the politics of language testing in immigration contexts (e.g. Hawthorne 1997; Shohamy 1997; Brindley and Wigglesworth 1998) and nicely complements a paper by Davies (1997c) documenting the history (and questionable ethics) of using English language tests as a means of barring undesirable aliens from entry to Australia.

It is fitting that the final paper in this section, by Eddie Williams, is about the status of English in developing countries, a subject which Alan Davies has been interested in (Davies 1968a, 1987, 1991c) since his early days as an English teacher in Kenya. In contrast with the other papers in this section, Williams’ article, Testimony from testees: The case against current language policies in sub-Saharan Africa, is concerned not with the social impact of language testing, but rather with what test outcomes tell us about the impact of language policy, in this case the policy of promoting English-medium instruction in primary schools in Southern Africa. While the rationale for the policy is one of modernisation and unification, Williams’ analysis of test results (relating to literacy in both English and local languages) suggests that this policy is limited in its effectiveness. His findings are contrasted with public perceptions regarding the value of English as a vehicle for social advancement.

Section Eight, The ethics of language testing, deals with some of the issues identified as critical by Davies (1997a and b) in a special issue of Language Testing (14,3) on test ethics and what he has called ‘right conduct’ of the professionals engaged in test development, use and validation. The section opens with a paper by Bernard Spolsky, Cheating language tests can be dangerous, which emphasises the fact that language test scores are at best ‘a chance approximation of the ability we hope to measure’. Like Davies, he believes that some form of testing is generally better than no testing at all but goes on to argue (using the history of the industrialisation of TOEFL by way of illustration) that testers’ energies have been misdirected: they have focused too much attention on the reduction of measurement error and too little on ‘the more urgent task of learning how to use flawed instruments fairly’.

Defining what we mean by fair test use is however a complex matter, as Liz Hamp-Lyons, in her paper, Ethics, fairness(es), and developments in language testing, points out. She reflects here on her early struggles with the issue of ethics in language testing in response to Alan Davies’ comments on her doctoral dissertation. She then goes on to describe the current difficulties she faces in ascribing a single meaning to the term fairness, given the large numbers of stakeholders involved in the testing enterprise and their sometimes conflicting views and needs. She issues a plea to language testers to pay greater attention to stakeholders’ perspectives and to assume greater responsibility for the impact of the instruments they devise.
Preface

Some of the scenarios proposed by Hamp-Lyons are quite at odds with traditional thinking about language testing and suggest the need for an alternative assessment paradigm. Lynch, in *The ethical potential of alternative language assessment*, talks of an alternative assessment culture involving an integrated view of teaching and assessment, in which students have a key role in making choices about how they are assessed and what counts as evidence in this process. He argues that alternative systems require different approaches to validity and proposes a theoretical framework for such approaches.

Section Nine, *Language testing and SLA*, takes the volume in a new direction, away from the consideration of tests and their social impact towards an investigation of their value as a tool for doing research into second language acquisition, including the measurement of vocabulary acquisition, grammatical development and ultimate attainment. The validity of empirical investigation in this area depends on satisfactory methods of measuring interlanguage development, as Davies (1990a), amongst others, has insisted.

The paper, *Quantitative evaluation of vocabulary: How it can be done and what it is good for*, by Batia Laufer, makes a case for using multiple quantitative measures of vocabulary acquisition both for practical purposes (e.g. to predict second language learners’ future academic performance) and in acquisitional research. She describes a number of such measures and illustrates their potential as research tools with data from recent research.

The issue of grammar testing, which featured strongly in the second section of this volume (see papers by Rea-Dickins and Clapham) re-emerges in a paper by Ellis, *Some thoughts on testing grammar: An SLA perspective*. Ellis questions the validity of indirect system-referenced practices in grammar testing (as exemplified by the TOEFL), because they measure only one type of grammatical ability. He offers what appears to be a response to the problems identified by Rea-Dickins and Clapham in the form of a set of provisional ‘specifications’ for instruments which are sensitive to both explicit and implicit grammatical knowledge and to sources of variation in interlanguage development.

Sorace and Robertson in their joint paper, *Measuring development and ultimate attainment in non-native grammars*, describe a measurement technique principally employed in psychophysics, known as magnitude estimation, which allows them to quantify judgements of linguistic acceptability on interval (rather than the traditional dichotomous or Likert-type) scales and therefore to use the full range of parametric statistics in analysing their findings. They demonstrate that magnitude estimation is especially suitable to the investigation of non-native competence, since it is sensitive to one of the main distinguishing features of these grammars: namely, indeterminacy and optionality at all stages of development.
The final section of the volume, *Beyond language testing*, contains two papers which fall well within the realm of Alan Davies’ interests and research activity but are not centrally about language testing. They nevertheless pose challenges to the language tester by identifying aspects of language behaviour which are not easily amenable to empirical investigation.

Selinker and Han’s paper, *Fossilisation: Moving the concept into empirical longitudinal study*, presents a state-of-the-art discussion of the poorly understood phenomenon of fossilisation (which was touched on in Davies and his colleagues’ early edited volume on Interlanguage (Davies, Howatt and Criper 1984)). The paper deals with both theoretical and definitional issues pertaining to the concept of fossilisation as well as the related concepts of stabilisation and multiple effects. The authors call for longitudinal studies as the only means of ascertaining that no change has occurred in an interlanguage form over time, and identify the need for context-sensitive language tests which can elucidate the complex nature of the fossilisation phenomena they identify. They argue as Davies has done (e.g. Davies 1998) for the importance of co-operation between language testers and SLA researchers.

John Maher’s paper, *The unbearable lightness of being a native speaker*, is a fitting conclusion to this volume. It takes up and elaborates, in a fanciful but scholarly fashion, the tantalisingly elusive concept of the native speaker which Davies has written about extensively (Davies 1991a and d, 1992b, 1994, 1995c). The notion of the native speaker encapsulates the kind of paradox (ideal versus real, universal versus particular) which has captured Davies’ imagination and which he has attempted to reconcile in much of his work in language testing and in applied linguistics more generally (see *The publications of Alan Davies* at the end of the volume). For testing, the challenge is to find a satisfactory compromise between principle and expediency (see Widdowson, Section Two) or, as Alan puts it, between uncertainty and explicitness.

*Our view is that through its important contribution to the fundamental linguistic tension between uncertainty and explicitness the central role of language testing in applied linguistics can be generally agreed.*

(Davies 1990a: 69)
Preface

References


Preface


Acknowledgements

A number of people, other than the editors and the contributors, worked very hard on the production of this volume. We would like to thank Samantha Green for tracking the contributors down, Natalie Stephens for collecting the abstracts and continuing the liaison work and Ruta Kanepe and Felicity Jensz for sending out reminders, circulating the incoming papers for comment and generally keeping the editors in line. Thanks also to Fiona Watson for her preliminary formatting of the volume.

The editors:

Catherine Elder, Annie Brown, Elisabeth Grove,
Noriko Iwashita, Tom Lumley, Kathryn Hill,
Tim McNamara, Kieran O’Loughlin
The University of Melbourne