The Australian Assessment of Communicative English Skills (ACCESS) is an English language testing instrument which is used around the world to assist the Commonwealth Department of Immigration and Multicultural Affairs (DIMA) in the migration selection process. It was developed by a consortium of Australian educational institutions under the aegis of NCELTR.

This volume gives an account of the process of test development during the period 1992 – 1994. It consists of two parts. The first part summarises the background and context of test development and administration. The second part presents a selection of research reports on major aspects of test construction and validation such as rater behaviour, the relationship between language skills and item difficulty and differences between direct and semi-direct oral tests.

The book will be of particular interest to language test developers, curriculum specialists, linguists and teachers throughout the world who have an interest in testing and assessment issues.

National Centre for English Language Teaching and Research
MACQUARIE UNIVERSITY SYDNEY AUSTRALIA
access:
Issues in language test
design and delivery

edited by
Geoff Brindley and Gillian Wigglesworth

National Centre for English Language Teaching and Research
Macquarie University, Sydney, NSW 2109
access: Issues in language test design and delivery

Published and distributed by the
National Centre for English Language
Teaching and Research
Macquarie University
NSW 2109

ISSN 1035 6487.
ISBN 1 86408 084 1.
I. Brindley, G.P. (Geoff P.). II Wigglesworth, Gillian. III. National Centre for English Language Teaching and Research (Australia). IV. Title: Australian Assessment of Communicative English Skills. (Series: Research series (National Centre for English Language Teaching and Research (Australia)); 9).
428.2407

MACQUARIE UNIVERSITY

© Macquarie University 1997

The National Centre for English Language Teaching and Research (NCELTR) is a Commonwealth Government-funded Key Centre of Research and Teaching established at Macquarie University in 1988. The National Centre forms part of the Linguistics discipline within the School of English, Linguistics and Media at Macquarie University.

Copyright
This book is sold subject to the conditions that it shall not, by way of trade or otherwise, be lent, resold, hired out, or otherwise circulated without the publisher’s prior consent in any form of binding or cover other than that in which it is published and without a similar condition including this condition being imposed on the subsequent purchaser.

All rights reserved. No parts of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the publisher.

NCELTR Research Series
Series Editor: Geoff Brindley

This series contains research reports on theoretical and empirical studies of significance to all those involved in the teaching of English as a second language to adults.

Design and layout: Stephanie Vaughan

Contents

Introduction
Geoff Brindley and Gillian Wigglesworth 1

Chapter 1
English language testing and immigration policy
Lesleyanne Hawthorne 9

Chapter 2
Issues in test design and delivery
Geoff Brindley, Sue Hood, Colin McNaught and Gillian Wigglesworth 31

Chapter 3
Investigating second language listening ability: listening skills and item difficulty
Geoff Brindley 65

Chapter 4
Negotiated discourse and interlanguage accent effects on a second language listening test
Steven Ross and John Langille 87

Chapter 5
Test-taker performance on direct and semi-direct versions of the oral interaction module
Kieran O’Loughlin 117

Chapter 6
The role of questionnaire feedback in the validation of the oral interaction module
Kathryn Hill 147
Chapter 7
Approaches to the evaluation of interviewer behaviour in oral tests
Janne Morton, Gillian Wigglesworth and Donna Williams 175

Chapter 8
A generalisability theory study of ratings and test design in the oral interaction and writing modules
Tim McNamara and Brian Lynch 197

Chapter 9
Text type and rater decision-making in the writing module
Susan Delaruelle 215
Introduction

Geoff Brindley and Gillian Wigglesworth

In 1992 Australia became one of the first countries in the world to include English language ability as a factor in its immigration selection procedures. To provide information on applicants’ proficiency, the Commonwealth Department of Immigration and Ethnic Affairs (DIEA) commissioned a comprehensive test of listening, oral interaction, reading and writing. The development and administration of the test, which became known as the Australian Assessment of Communicative English Skills (access), was initially carried out between 1992 and 1994 by a consortium of institutions under the aegis of the National Centre for English Language Teaching and Research (NCELTR) at Macquarie University. From 1995 responsibility for ongoing development and delivery of the test has been in the hands of the International Development Program of Australian Colleges and Universities (IDP) in partnership with the National Languages and Literacy Institute of Australia (NLLIA), Language Testing and Curriculum Centre, at Griffith University.

In conjunction with the development of the access test, DIEA funded an ongoing program of research aimed at investigating the validity, reliability and practicality of the test. It was intended that the outcomes of this research would be continuously fed back into the decision-making process and lead to improvements in test design and administration. This volume brings together a selection of the studies which emerged from this research program during the period 1992–1994, along with two additional papers describing the context in which test development took place.

The volume is structured so that each contribution may be read independently, although the chapters are conceptually linked through their focus on different aspects of the access test and, in some cases, through their use of the same analytical tools and techniques. The first two chapters provide an overview of the policy background and test development process. The remaining chapters present the outcomes of research studies of the listening, oral interaction and writing modules of the access test. Collectively, they illustrate the issues and dilemmas which arise in test development and validation and suggest how these can be investigated using both qualitative and quantitative methods. Some of the key questions addressed by contributors are:
• What is the relationship between the listening skills described in test specification and actual test performance?
• What is the effect of including interactionally modified discourse in a listening test?
• How comparable are learners’ performances in direct and tape-mediated oral tests? How valid do each of these formats appear to test takers?
• What is the influence of rater behaviour on oral scores?
• How many tasks and how many raters are necessary to ensure acceptable levels of reliability in tests of written and spoken production?
• Do different writing tasks require specific rating criteria or can generic scales be used across a range of tasks?

We hope that the results of these studies will not only provide a basis for improving methods of test construction and validation, but will also, through the issues and questions they raise, help to stimulate further research by language test developers.

The opening chapter of this volume, by Lesleyanne Hawthorne, outlines the political and economic background to the development of the access: test. Hawthorne traces the recent history of skilled migration and identifies some of the problems confronting skilled migrants with limited English proficiency. She then traces the changing significance of English language ability as a factor in the immigration selection process and demonstrates how the greater emphasis on English language skills in the immigration policy created a need for a formal test to replace past ad hoc assessment practices. Hawthorne draws on interviews with the Chair of the test development committee to outline the process of development and delivery of access: The ways in which mandatory English language testing has been used as a policy tool by DIEA are also discussed.

The author concludes by identifying potential problem areas arising from the introduction of mandatory English testing, noting the need to ensure that candidates are not unfairly disadvantaged by their ethnic background, the fee regime, or by the setting of unnecessarily high standards.

Chapter 2 by Geoff Brindley, Sue Hood, Colin McNaught and Gillian Wigglesworth provides a reflective account of the test development process from the perspective of the test development team. They begin by describing the practical constraints under which the test was developed and show how the eventual shape of the test was determined by the need to find a balance between validity, reliability and feasibility. The authors then take the reader through the process by which the test specifications for each subtest were translated into test items and tasks, providing an overview of the principles that were used to define the content of the tests, to select tasks, and to decide on the types of item formats that were used. They also discuss ways in which information from statistical analysis of test data was used to modify and improve the tests.

In their account of the test development process, these authors outline some of the knotty dilemmas which were faced when constraints of time and resources had to be balanced against the need to design a valid and reliable instrument which reflected current understanding of language proficiency. The chapter concludes by affirming the need for test developers to undertake ongoing research and validation as a basis for continuous improvement.

Chapter 3, by Geoff Brindley, focuses on the question of listening skill difficulty in the access: listening module. Brindley begins by describing the wide variety of factors which may affect second language listening ability, and identifies a number of specific factors which have been shown by research to be key determinants of difficulty. He points out, however, that studies of listening text and task difficulty have yielded contradictory results and that clear evidence for clearly definable sub-skills or difficulty hierarchies is still lacking.

Brindley then reports the results of a study into the access: listening module which aimed to investigate the relationship between listening skills and item difficulty. He finds a weak relationship between the judges’ ratings and actual item difficulties based on candidates’ results, and little agreement on the assignment of listening skills to particular items. He attributes this lack of agreement to the fact that any item may draw on a variety of skills simultaneously, thus making it almost impossible to relate particular skills to particular items. He also sound a note of caution regarding the practice of asking native speakers to introspect on the difficulty of items for language learners and calls for more research into the test-taking processes and strategies used by the actual population for whom the test is designed. Brindley concludes by suggesting that although skills taxonomies probably cannot be used as a prescriptive blueprint for item design, they may help listening-test developers to check the extent to which different skills are being sampled and, thus, help to improve the test’s content validity.
Chapter 4 by Steven Ross and John Langille also focuses on factors affecting listening comprehension test difficulty. They report the results of a study involving 588 Japanese and Finnish learners of English which set out to investigate the impact of negotiated interaction on comprehension and specifically to determine to what degree the inclusion of interactionally modified discourse in test dialogues affects the comprehensibility of non-native accents. In a carefully designed study using internal replication they examine differences between scores in two parallel forms of a test containing native speaker/non-native speaker dialogues, one with negotiated interaction centring on difficult novel lexical items, the other with no negotiation. They also examine differences in the comprehensibility of specific propositions and lexical items spoken by the non-native speakers used in the experiment compared to those uttered by the native speakers.

Their hypotheses that negotiated dialogues will result in greater comprehension of propositions and lexis are largely supported, although the impact of negotiation is mediated by language proficiency level. In the accent study they find no main effect attributable to accent differences between the native and non-native speakers.

In considering the implications of the study for test design, the researchers point out that the presence of accommodative negotiation in discourse mirrors authentic interaction in multicultural contexts. For this reason they argue that there are grounds for including negotiated episodes involving non-native speakers in test dialogues. If this is done, the degree of difficulty of a test passage can be manipulated by controlling the amount of accommodative interaction as well as the degree of accentedness of the discourse.

The adoption of Ross and Langille’s suggestions could lead to specifications which would be quite different from those which underpin current listening tests. Their findings on the effects of redundancy are a timely reminder of the gap which often exists between the idealised kind of native speaker discourse which figures in language tests and that which actually occurs in multilingual societies.

If tests of ‘English as an international language’ are to have any claim to reflect the figures in language tests and that which actually occurs in multilingual societies, the scope of propositions and lexis are largely supported, although the impact of negotiation is mediated by language proficiency level. In the accent study they find no main effect attributable to accent differences between the native and non-native speakers.

In considering the implications of the study for test design, the researchers point out that the presence of accommodative negotiation in discourse mirrors authentic interaction in multicultural contexts. For this reason they argue that there are grounds for including negotiated episodes involving non-native speakers in test dialogues. If this is done, the degree of difficulty of a test passage can be manipulated by controlling the amount of accommodative interaction as well as the degree of accentedness of the discourse.

The adoption of Ross and Langille’s suggestions could lead to specifications which would be quite different from those which underpin current listening tests. Their findings on the effects of redundancy are a timely reminder of the gap which often exists between the idealised kind of native speaker discourse which figures in language tests and that which actually occurs in multilingual societies. If tests of ‘English as an international language’ are to have any claim to reflect the figures in language tests and that which actually occurs in multilingual societies, the scope of propositions and lexis are largely supported, although the impact of negotiation is mediated by language proficiency level. In the accent study they find no main effect attributable to accent differences between the native and non-native speakers.

In considering the implications of the study for test design, the researchers point out that the presence of accommodative negotiation in discourse mirrors authentic interaction in multicultural contexts. For this reason they argue that there are grounds for including negotiated episodes involving non-native speakers in test dialogues. If this is done, the degree of difficulty of a test passage can be manipulated by controlling the amount of accommodative interaction as well as the degree of accentedness of the discourse.

The focus of Chapter 6, by Kathryn Hill, is also the test takers themselves. The author relates test-taker feedback to issues of test validity and comparability, arguing that test takers may provide insights into test design and that their perceptions may also contribute to the validity of a test, specifically its face validity. Hill examines questionnaires which were completed by test takers subsequent to the trialling of an access: oral interaction module in December 1992. Test takers had undertaken both the live and the tape format of the test and so were in an ideal position to make comparisons between the two — a situation which never arises in a live administration. Due to the broad range of test takers who participated in the trial, Hill is able to compare responses from different groups of test takers. She investigates whether language background (European versus Asian), gender, or professional status (student versus professional) influence attitudes to the test.

Hill finds favourable reactions to both test formats, with an overwhelming preference, however, for the live format — although this response was not uniform
across different groups. She also identifies a number of aspects of the tests which were problematic for test takers. In her comparison of test performance and reactions to the test across groups, Hill finds no major differences in performance between professional and student candidates, although the former were more positive in their reactions. Nor was there a clear effect for language background amongst the target group for the test. Hill suggests, however, that her failure to find clear differences in performance between the target and non-target groups need not be seen as a threat to test validity since access: is not intended to be a specific purpose vocational test. Encouragingly, the test's high face validity with the professional group would seem to indicate that it is perceived by the population for whom it is intended as appropriate for its purposes.

In Chapter 7 Janne Morton, Gillian Wigglesworth and Donna Williams address the question of interviewer variability using quantitative and qualitative techniques to investigate differences between interviewers who administer the oral interaction test. During the rating process raters are routinely asked to complete open-ended questionnaires on the performance of the interviewers. For the purposes of this study raters were also asked to provide a rating for each interviewer and from this analysis a group of good raters and a group of poor raters were identified. Statistical analysis, in line with previous findings (e.g. Lumley and McNamara 1995), revealed that raters tended to compensate for interviewers they perceived as being poor, and that non-native interviewers were more frequently classified as poor — the latter finding providing empirical support for the decision to use only native-speaker interviewers in the access: test.

The researchers then use the qualitative comments of the raters to identify features of good interviewer behaviour. In addition, they compare the discourse of the structured tasks in the test with the discourse of unstructured task (the role play) and find that the structured tasks encourage greater uniformity of rater behaviour, thus diminishing the likelihood of candidates’ scores being adversely affected by poor interviewing techniques. Thus, while the decision to use such a test was taken because it needed to be comparable with the tape-based version, this finding provides independent evidence to support this decision.

In Chapter 8, by Tim McNamara and Brian Lynch, the important question of measurement error is addressed. While an element of error is present in any measurement, in the testing situation it is essential to maximally reduce this. Although increasing the number of raters assessing the test reduces error, this measure must be balanced against practicality and the additional expense involved. Similarly, the number of tasks administered in any test must be balanced against practicality, expense and the ability of the candidate to maintain concentration and attention over the period of time required to complete the tasks. Thus it is important to identify the optimum number of tasks, and the optimum number of ratings required to maximally reduce error and ensure candidates are being fairly treated. McNamara and Lynch use Generalisability Theory to investigate this question for both the oral interaction and the writing modules of the access: test. They conclude that consideration could be given to a reduction in the number of tasks included in these two modules, and that two ratings for each performance are essential, with a third rating being provided for borderline cases, or cases where the two raters disagree significantly.

Chapter 9, by Susan Delaruelle, uses rater talk-aloud protocols to investigate the criteria used to assess writing performance. The study was designed to investigate the validity of using the same set of rating descriptors for all three tasks in the access: writing module. Two different text types — one interpersonal, the other expository — were identified on the basis of their functional and social purpose and six naive raters were asked to rate each text-type on a six-point rating scale for which they did not have descriptors. Three of the raters were highly experienced ESL teachers while the other three, who were also ESL teachers, had considerably less experience. Delaruelle analyses the protocols along two main dimensions — frequency of mention of different criteria and thematic prominence (i.e., category first mentioned). From her analysis she is able to identify a number of salient criteria according to which the raters assess both text types. On the whole she finds that the criteria used are very similar across the different text types and concludes that this finding provides evidence to support the use of the access: writing rating descriptors in their present form.

Notes

1. Now the Department of Immigration and Multicultural Affairs (DIMA).
2. Now Language Australia.

Reference

English language testing and immigration policy

Lesleyanne Hawthorne

Skilled migration and English language testing

In 1992, with Australia in deep recession and pressure mounting for ongoing reduction of immigration, evidence was already beginning to emerge of the disadvantages non-English-speaking background (NESB) workers faced in the labour market (Foster et al 1991; ABS 1992; Birrell 1992; DIEA 1994a; Wooden 1994).

Many of these workers were highly skilled professionals who had been selected for migration to Australia in the late 1980s, during a period of unprecedented expansion of the skilled migration program following publication of the report of the Committee to Advise on Australia’s Immigration Programs (CAAIP 1988).

As a result of this expansion, 70,943 skilled migrants and their dependents — representing 58.3 per cent of the total migrant intake — arrived in Australian during the 1990–91 period. A comparison of the statistics for engineers migrating to Australia during this period and during the past twenty years gives some indication of the scale of these arrivals. Between 1987 and 1993, 18,581 engineers migrated, compared with an average of 817 per year over the previous twenty years (Rice 1992).

A growing number of these skilled arrivals were from non-English speaking backgrounds — a reflection of the fact that by 1990 NESB migrants possessing tertiary qualifications outnumbered two to one English-speaking background (ESB) migrants with similar qualifications. In addition, and in marked contrast to earlier Australian selection practice, NESB migrants were drawn from an increasing number of source countries (Hawthorne 1994a). By the period 1990–91, for instance, the proportion of UK engineers had shrunk to 10 per cent of the total intake of engineers while migrants from the rest of Europe amounted only to a further 8 per cent. Skilled engineering migration had become a largely
Asian phenomenon, with North-East Asia accounting for 31 per cent of arrivals, South Asia for 21 per cent, and South-East Asia for 16 per cent. The other 14 per cent includes migrants from a variety of countries, including Central and South America and the Middle East.

Table 1: Percentage of settlers arriving in Australia (top ten countries of birth)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Country of birth %</td>
<td>Country of birth %</td>
<td>Country of birth %</td>
</tr>
<tr>
<td>UK &amp; Ireland 48.5</td>
<td>UK &amp; Ireland 41.3</td>
<td>United Kingdom 13.5</td>
</tr>
<tr>
<td>Greece 12.2</td>
<td>Yugoslavia 5.6</td>
<td>Hong Kong 12.0</td>
</tr>
<tr>
<td>Italy 10.5</td>
<td>New Zealand 3.6</td>
<td>Vietnam 8.9</td>
</tr>
<tr>
<td>Yugoslavia 4.5</td>
<td>Greece 3.4</td>
<td>New Zealand 6.7</td>
</tr>
<tr>
<td>Malta 3.7</td>
<td>USA 2.9</td>
<td>Philippines 5.5</td>
</tr>
<tr>
<td>Germany 2.9</td>
<td>Turkey 2.8</td>
<td>India 5.2</td>
</tr>
<tr>
<td>Netherlands 1.9</td>
<td>Lebanon 2.8</td>
<td>China 3.2</td>
</tr>
<tr>
<td>Egypt 1.6</td>
<td>Italy 2.6</td>
<td>Taiwan 3.0</td>
</tr>
<tr>
<td>New Zealand 1.4</td>
<td>India 2.2</td>
<td>Malaysia 2.9</td>
</tr>
<tr>
<td>USA 1.1</td>
<td>Malta 2.0</td>
<td>Sri Lanka 2.6</td>
</tr>
</tbody>
</table>

The migration of these skilled NESB workers was accompanied by mounting concern within key Australian professions, not only about the scale of intakes but also about the professional calibre of those selected — an issue which had rarely been raised in periods dominated by ESB skilled migration. Although such concerns were in some cases clearly motivated by labour market protectionism (Kunz 1975), many professional bodies argued — with some justification — that English language ability and employability were closely linked. According to Martin Thomas, President of the Institution of Engineers, Australia (IEAust) (Thomas 1991:46):

“... Australia, at least in the immediate term, is accepting too many migrant engineers. While meeting education requirements sufficient for graduate membership ... and hence gaining immigration points, many are not adequately skilled, whether by language capability, postgraduate experience or current competence, in the most needed engineering skills in Australia to find a job ... Some members of IEAust have suggested, and I share their views, that in addition to [qualifications recognition] attention should be paid also to prospective immigrants' professional experience and English language skills before immigration rather than after arrival.”

In the view of Ralph Blacket, Chairman in 1990 of the Australian Medical Council's Examination Committee (Blacket 1990: 129):

“... all English speaking developed countries take the view that foreign medical graduates who choose to emigrate must demonstrate in objective testing a good level of proficiency in English as well as the level of professional competence expected of graduates in their chosen country.

These criticisms of migration policy highlighted two significant issues. The first was the question of mandatory English language testing in occupations which required vocational levels of English, and the second was need for an objective test taken by NESB professionals, either as a pre-condition for their migration or within Australia prior to professional registration.

Although by the late 1980s pre-migration English language testing was compulsory for skilled NESB applicants in theory, English language ability was in fact waived as a condition of entry for Concessional Family applicants, and was not systematically applied to Independent applicants — the majority of whom self-assessed their level of English ability in the absence of any independent testing instrument (Hawthorne 1994a). Many skilled migrants were thus reaching Australia with minimal competence in English and, in the eyes of employers and professional groups, this was a major problem creating serious barriers to their immediate professional employment.

The evolution of English language testing: 1980 to 1992

Although the concept of vocational testing was still in its infancy in Australia in 1980 and there were as yet no standardised English language tests for administration at overseas posts, English language competence was nevertheless already part of the points-weighted Numerical Migrant Assessment System (NUMAS) devised by the (then) Department of Immigration and Ethnic Affairs (DIEA) to screen candidates applying for migration.

Introduced in 1979, this points system was designed to select skilled migrants on the basis of their ‘economic and settlement prospects in Australia’ and their ‘personal capacity to settle successfully’ (DIEA 1981: 3). It was the proclaimed intent of DIEA that this assessment be impartial and should employ ‘uniform standards through the same procedures used around the world’. Under NUMAS, English language competence could gain the applicant up to 10 points towards the necessary number of points of acceptance as a migrant. Up to three of these
points were awarded in relation to economic capacity, two for ‘ability to communicate in a proposed occupation’, and five for personal and settlement potential (since knowledge of English was deemed important to a migrant’s access to, and use of, the various services available in Australia).

This ten-point maximum for English actually exceeded the number awarded for critical economic attributes such as recognised skills. As it transpired, there was some justification for this weighting given emerging evidence linking Australian employment outcomes with facility in English (Bureau of Labour Market Research 1986). Despite the DIEA commitment to impartiality, however, assessment of skilled applicants in relation to English competence and many other factors was largely conducted at the discretion of individual migration officers. No formal English language proficiency tests were used at overseas posts and no training in language assessment was provided to DIEA staff. It is not surprising that officers made highly subjective assessments and generated very variable results for different individuals within and across posts.

In 1983, explicit weighting for English temporarily vanished from the points-tested system, but English language skills continued to be assessed in a similarly subjective fashion. Within the new Migration Selection Program, migration officers were required to rank applicants’ perceived settlement prospects — from ‘settlement risk’ to ‘outstanding’ — across a five-point scale. This was frequently done on the basis of a brief, monolingual interview (DIEA 1983). English language skills were included either implicitly or explicitly along with many other variables. Judgments were required not merely on applicants’ current ability, but also on their perceived future settlement aptitude and learning potential.

This covert language assessment persisted through the mid-1980s. English language ability only reappeared as a discrete (though not compulsory) item after the substantial revamp of Australia’s points system following the 1988 publication of the influential Report of the Committee to Advise on Australia’s Immigration Policies (CAAIP) which emphasised selection of migrants with proven competence in English. In the Committee’s view, such a selection was a matter of common sense and necessity in order to ensure that migrant workers would be able to upgrade their skills as required in response to constant technological development (CAAIP 1988).

Publication of the CAAIP report was followed by a four-year surge in immigration, characterised by a heightened emphasis on skills. As part of this process, English language ability was theoretically reinstated as a significant criterion in Independent applicants’ points testing. By April 1989 a maximum of fifteen points were given for language ‘proficiency’, a score of five meaning the applicant had ‘only limited English ability’ (DIEA 1989).

This re-emphasis on English language skills, however, failed to lead to the decisive screening of English language ability advocated by CAAIP. There was still no assessment requirement for Concessional Family applicants and even for Independent category applicants the points allotted for English skills were not critical to their selection if they were of prime workforce age and had tertiary qualifications. In any case, with the shortage of staff at Australian overseas posts principal applicants could take a gamble and award themselves the maximum fifteen points in their self-assessment of language ability, since it was rare for DIEA to demand any verification. According to a Yugoslav engineer who migrated in 1990 (Hawthorne 1994a: 41):

I saw a lot of people who achieved the required score by giving themselves fifteen points. There was absolutely no checking! … I must tell you that I have a friend who told me how to pass the examination (if one should be held) … He told me that he was informed by his friends of eight or ten standard questions. He learned them by heart, and he learned the answers by heart. Even though he didn’t know any English at this time, he passed the interview! It is really amazing how the Australian Government allows this to happen.

In the wake of the CAAIP report, the explosion of migration enquiries from all over the world, coupled with resource limitations in overseas offices, meant that the assessment of prospective skilled migrants was carried out largely on paper. In January 1994 there were 114 DIEA officers, supported by local staff, at overseas posts. A year earlier, in 1992–93, similar levels of staff had been required to deal with 2,028,000 temporary entry applications, in addition to permanent entry applications from 109,000 people. According to DIEA, in cases where levels of English could make a critical difference to acceptance or rejection an attempt was made to test language skills using informal procedures. However, given the lack of accredited tests, the results remained highly variable.

This anomaly inevitably saw the migration to Australia of many Independent applicants as well as Concessional Family applicants with minimal levels of English — a situation which increasingly concerned DIEA. Many NESB migrants
required a considerable period of post-arrival English language instruction before they could hope to find work and it was not uncommon for skilled migrants to spend twelve to eighteen months learning English, only to commence job-hunting handicapped by acute career gaps. Increasingly, feedback from both employer and professional groups confirmed that advanced English language competence was essential for professional employment. This typically meant that ‘you do not get a foot in the door if you do not clearly speak English’ (Hawthorne 1994a). Though employers might have tolerated poor language skills in times of comparative labour market demand, this pattern was reversed in a period of over-supply — with very serious consequences for the job prospects of incoming professionals.

Changes to immigration policies heralded by mandatory English-language testing: 1992

In July 1992, DIEA took steps to underline the importance of English language skills in the migration program by introducing a mandatory English-language requirement for the following three major groups of applicants:

- skilled NESB migrants qualified in key professional fields, henceforth to be known as ‘Occupations Requiring English’ (OREs);
- independent category principal applicants who were to be tested in order to determine the number of points (out of a maximum of 20) which could be earned for English language competence under the revamped points system;
- other independent and concessional applicants, including adult family members of principal applicants who were to be tested to determine whether they would be entitled to English as a Second Language tuition in Australia.

The introduction of mandatory English language testing represented a radical departure from past Australian practice in a number of ways. Firstly, it meant that immigration points were now directly linked to levels of vocational English language competence. Since the points were allocated by DIEA rather than the test developers, the assessment of English became a direct and variable instrument for the control of skilled immigration intakes. Secondly, ‘vocational competence’ in English became a mandatory criterion for people whose jobs appeared on the ‘Occupations Requiring English’ list which embraced an extremely comprehensive range of professions and trades related to health, engineering, public safety, education, and policing (Commonwealth of Australia 1993) (see Appendix). From this time ORE Concessional as well as Independent principal applicants who failed to demonstrate vocational level English would not be permitted to migrate to Australia, however many points they gained on other criteria. Vocational English was defined by DIEA (1993a: 26–27) as the ability:

... to speak and understand English well and with sufficient vocabulary to participate fully, both formally and informally, in a wide range of work and social situations, with a fairly high degree of fluency and precision.

Thirdly, the new emphasis on English language skills offered a substantial competitive advantage to candidates from English-speaking backgrounds or from countries with strong traditions of English-medium instruction (such as Hong Kong or India). A maximum twenty points could now be allocated to Independent applicants who possessed ‘vocational competence in English’. An Independent principal applicant in 1994, for example, could gain the 100 points required for acceptance as a migrant by possessing recognised qualifications supported by three years experience (70 points), and twenty through high level competence in English, thus scoring ninety points before the awarding of any additional points for age.

Finally, off-shore English language testing was to be used not only as part of the migration selection process but also as a means of determining the extent of applicants’ need for English language instruction on-arrival. Any Independent or Concessional applicant (principal applicant as well as adult dependents) defined as requiring further ESL instruction would be required to pay, as a pre-condition of application processing, significant up-front fees to cover tuition. The fees ranged from $AUD4080 to $AUD1020 per person and are explained in the following words in a DIEA publication (1993b):

*The charge is payable for each applicant aged 18 years or over who is assessed as not having functional English. If you are in one of the categories required to pay the charge … you must pay the charge before your visa is granted … On arrival in Australia you will then be provided with up to 510 hours of English tuition, or as many hours as it takes for you to achieve this level of proficiency, whichever comes first.*

An addition in a later DIEA publication (1993c) stated:

*If the charge is not paid the visa will not* (repeat not) *be granted.*
This fee requirement was part of the introduction of ‘user pays’ principles into a revamped Adult Migrant English Program and represented a major departure from past DIEA practice. The introduction of the education charge, however, clearly constituted a potential pre-migration deterrent to poorer NESB applicants.

The development of access: 1992–94

The National Centre for English Language Teaching and Research (NCELTR) at Macquarie University was commissioned to develop the new test which was to become known as the Australian Assessment of Communicative English Skills (access). When Professor Chris Candlin, Executive Director of NCELTR, was approached by DIEA to develop the test in August 1992, he foresaw significant challenges:

[The] administrative challenges ... in setting up such a kind of worldwide system [over a very short period] were clearly considerable. Secondly we were concerned about issues of equity. Thirdly we were concerned about cost.6

To a degree, Candlin’s concerns were fuelled by awareness that despite the evidence of NESB immigrants’ comparative labour market disadvantage (Wooden 1994), there was in fact no clear cut relationship between English language ability at time of entry and successful resettlement. In fact, the history of post-World War II migration to Australia is studded with stories of NESB migrants who arrived without English, yet rose to positions of wealth and influence. In addition, the extent to which immigrants learn language successfully is heavily influenced by, among other things, motivational factors, educational background, age, learning skills and cultural knowledge.

However, a language test could clearly not assess such factors or screen for learner potential and it therefore became very important to develop a system which provided a fair and accurate assessment of a candidate’s language skills at the time of application. This would involve a substantial investment on the part of the government if the test were to meet the required standards. Describing the context in which the development of the test was initially undertaken, Candlin comments:

I think that right from the beginning our argument to the Department was that the degree of investment you make in reliability and validity in tests is obviously commensurate with the importance for individuals of the test, and that in this case that importance was very high ... We have had a Department representative at our test development meetings, and there has always been a very strong argument in favour of having double rating, of rater training, of interviewer training, and all of this has been accepted by the Department without demur ... We have had nothing but the strongest support from them, in terms of those matters that we felt have been necessary to ensure reliability and validity ... We (also) got a commitment from the Department from the beginning for a research agenda, which would contribute to the reliability of the test.

From the beginning, Candlin counselled against the use of existing tests, such as the International English Language Testing System (IELTS), except as an interim measure. The IELTS test was designed in the late 1980s and is used internationally to assess NESB students’ readiness to cope with tertiary study in English. Candlin had earlier been closely involved with IELTS for five years, including his role as Chair of the International Editing Committee. In his view as IELTS is a test of English for Academic Purposes, it had limited capacity to assess ‘vocational’ levels in the context of immigration to Australia. In addition, the test had largely been trialled with young Singaporean and Hong Kong students rather than on a representative cross-section of candidates worldwide and there was a possibility that the test might not be valid for other types of candidates. Until recently, regular recycling of IELTS test items had also been permitted — a practice with clear implications for test security.

Over the following twelve months a consortium headed by NCELTR and involving the National Languages and Literacy Institute of Australia (NLLIA) Language Testing and Research Centre (LTRC) at the University of Melbourne, the NLLIA Language Testing and Curriculum Centre (LTCC) at Griffith University (until 1994), and the Adult Migrant English Service of New South Wales, assembled a team to develop and administer the access: test.

A number of steps were taken by the test developers to ensure that the standards of test development were commensurate with internationally accepted practice. These are described in Chapter 2. At the same time a program of research projects — many of which are described in this volume — was initiated as a basis for monitoring and improving the test. In locations where access: or IELTS were not yet available, DIEA Migration Officers were provided with enhanced training and guidelines to assess NESB candidates — an improvement over previous more ad hoc practice (Hawthorne 1994b).
The first overseas administration of access: was in April 1993 and by October 1994, access: was available in 27 locations, with seven further locations scheduled to come on-stream in the next four months. A target of late 1995 was set for global availability. Examination centres were sought which could offer a secure test environment (such as British Council offices, Australian Education Centres, IELTS outlets and established educational institutions).

In countries where language laboratories were frequently unavailable, or unreliable due to lack of technical maintenance, the access: consortium, drawing on NCELTR’s international network, activated a wide range of personal and professional contacts to secure appropriate venues. Though access: was offered on a user-pays basis, fees were scaled in line with specific location costs, taking into account local economic conditions. Candidates were guaranteed a six-week turnaround in results.

By 1994 an access: candidate information booklet had been prepared providing a description of the test supported by samples of a range of test items covering social and vocational contexts (NCELTR 1994). As it was designed to encourage candidate self-assessment, the handbook was written in English, at intermediate level, rather than in local languages. The handbook also outlined six possible levels of English language competence. An attainment of Level 6 in all four skills — speaking, listening, reading and writing — was originally required for both Occupations Requiring English and for Independents wishing to secure the maximum ESL points test score of twenty points. This level almost certainly matched Australian employer expectations concerning ‘vocational’ English (Hawthorne 1994a). The handbook described Level 6 in the following terms (NCELTR 1994: 7):

**Level Six**

You can read and understand a wide range of English texts easily and with good comprehension; you can write English appropriately and with quite a high degree of accuracy for a range of purposes; you can easily understand spoken English in a wide variety of situations; you can speak English appropriately and with quite a high degree of accuracy and fluency in most cases.

In May 1994, the requirement to attain Level 6 was substantially eased for OREs, who were now obliged to meet Level 5 standards instead. Level 5 required candidates to read ‘with reasonably good comprehension’, write ‘well enough to communicate effectively for most purposes’, understand ‘quite competently’, and speak ‘fairly fluently and accurately’ (NCELTR 1994).

In addition, ORE applicants were now required to achieve ‘vocational’ English levels (ie Level 5) in only three rather than four of the language skills — a concession which allowed for one skill area to be lower than the others (DIEA 1994a). NCELTR had advised a degree of latitude on this, given that otherwise excellent applicants might be judged to have less advanced skills in relation to speaking, for instance, due primarily to lack of opportunities for practice. DIEA later justified the decision in the following way (DIEA 1994b):

*The changes to the measurement of English language points recognise both that English language competency needs to be measured across a wide range of skills, and that the top score previously required was generally beyond occupational requirements.*

**Mandatory ESL testing of skilled migrants: the issues**

The introduction of mandatory English language testing by DIEA is aimed at alerting skilled NESB applicants in advance to the link between English language skills and employment outcomes, thus ensuring that incoming migrants more immediately meet the expectations of Australian employers (DIEA 1994b). Furthermore, it represents a way of achieving substantial saving for DIEA in terms of post-arrival English language instruction, since ORE professionals are in principle defined as ineligible for government-funded language programs, and those who do not demonstrate a Functional level of English in the access: test must pay their tuition costs prior to emigration.

At present, however, there remain four key policy issues to monitor in relation to the mandatory testing of English.

**The potential of access: levels to control immigration numbers**

The first issue relates to ‘the politicisation of English’ — the use of testing as a significant means of controlling overall numbers within Australia’s skilled migration program (Hawthorne 1994b).

Prior to the introduction of mandatory English language testing, levels of skilled migrants were high, with the majority of those arriving from NESB source countries. However, by 1993–94 numbers had dramatically reduced, for example
in the Independent category immigration fell from 30,160 in 1991–92 to 9171
in 1993–94, and in the Concessional Family category from 21,325 in 1991–92
to 8107 in 1993–94 (BIPR 1994).

According to a DIEA spokesperson, this reduction occurred across the immi-
grant program and could not be directly attributable to the introduction of
testing. Rather it was likely to be a reflection of reduced international demand
for migration to Australia at a time of recession, reinforced by the raising in 1992
of the number of points required for immigration (100 for Concessional and 110
for Independent applicants), together with the introduction of a cluster of fees
for ESL testing, on-shore ESL training, qualifications and health assessment.

Notwithstanding this, it seems likely that mandatory testing has had at least
some degree of impact since its progressive introduction from mid-1992, partic-
ularly given that both IELTS and access: initially had very high target scores.
Moreover, the potential had now been built into the system to use English
language skills as a ‘drawbridge’ within the skilled migration program.

The consortium developing and administering the access: test advised DIEA on
the characteristics of different levels of ability, but it was the Department who
made the ultimate decisions on the standard required for immigration. As
Candlin comments:

What we’ve said … is that the values the Department chooses to put
upon the results of the test in terms of migration points is a decision
that the Department makes — not us. We can say to you in respect of
levels … where we wish to place the person. It’s for you, as user of that
information, to decide what you’re going to do with it from a policy
point of view … Migration Planning were the people who wanted to
keep this flexibility, they wanted to be able to raise or lower the [level],
so to speak, in response to other factors.

It is significant to note here that very high pass levels were specified for both
access: (Level 6) and IELTS (Band 7) while Australia was in recession and wished
to substantially cut skilled migrant intakes (BIPR 1994). In 1994, with the eco-
nomic recovery and a cautious rise in the migration program, language require-
ments were significantly eased (Hawthorne 1994b). The easing also reportedly
occurred because migration officers felt otherwise suitable applicants were being
excluded by an unduly high pass level. Recently published data confirms that
68 per cent of principal Independent applicants gained the maximum 20 points

for English in 1993–94, with a further 7 per cent gaining 15. This compares with
a rise to 75 per cent and 8 per cent respectively a year later (Birrell 1995).

The potential for bias

Although a number of measures have to date been taken to try to ensure that
access: operates without bias against particular groups, applicants from English
speaking backgrounds are clearly advantaged in terms of pre-migration language
testing given that they are exempted from taking the test, are awarded the maxi-
mum 20 points for English language ability and automatically pass in the cate-
gory ‘Occupations Requiring English’. Interestingly, however, this issue has
aroused, to date, minimal response from Australian ethnic communities —
perhaps due to NESB professionals’ disproportionate unemployment during
the period of recession.

Certain NESB groups are also advantaged within this process, particularly appli-
cants from former Commonwealth countries with English-medium education
systems such as India, Sri Lanka, Malaysia and Hong Kong, and also west Euro-
peans from countries with a strong tradition of English language teaching. In
terms of numbers, the former will certainly predominate over the latter, given the
variation in regional levels of demand for migration. Many of these applicants
are in fact exempt from language testing — for example those who have ‘under-
taken … higher education in an institution where ALL instruction was con-
ducted in English. The qualification must require at least three years full time
study’. (DIEA 1994a)

Other NESB groups (for example east Europeans) are likely to fare comparatively
poorly. NCELTR was aware of this and brought the matter to the attention of
the Department as Candlin remarked:

[This] is a matter that is very clear to us, that if you come from Hong
Kong your chances of doing well on access: are infinitely better than if
you come from Kazakhstan. And we said to them ‘Look, the evidence
from the first round of access: shows that people coming through the
Moscow centre are doing proportionately less well than people from
other centres. We may need to consider this.

Candlin acknowledges that any pre-migration procedures — including health
checks and qualifications recognition requirements — will distort selection pro-
cesses to some degree. At the same time, however, the introduction of access:
clearly has the potential to ‘skew’ selection outcomes by ethnicity, given the different exposure of candidates to English language instruction.

A related issue concerns access and equity — the imposition on NESB applicants of pre-migration fees to cover language testing and (if necessary) the substantial costs of subsequent Adult Migrant English Program (AMEP) instruction ($2040 for Independent principal applicants, spouses and each adult dependent) (DIEA 1994a). ESB applicants are exempt from such fees, while applicants from relatively affluent regions are not significantly disadvantaged. By contrast, the fees may constitute a major barrier for less materially advantaged ethnic groups — for example, non-Humanitarian applicants from the former East European bloc, who may have transitional problems with English, but who otherwise represent excellent skilled migrant potential. (However, Humanitarian entrants, even in the Occupations Requiring English category, are exempt from the education charge.)

DIEA, along with other government departments, has defended the introduction of a partial user-pays system, arguing it has in fact coincided with growth in AMEP client participation (Storer 1995). Moreover, as the English Language Intensive Courses for Overseas Students (ELICOS) experiment has shown, intending applicants from even relatively poor nations such as the People’s Republic of China have proven able to find sums of $5000 or more to study English overseas.

There remains a clear necessity, however, to monitor access and equity issues in relation to ESL off-shore fees and testing.

**English language testing and the composition of the skilled migrant intake**

A further critical issue relates to the way in which mandatory English language testing may influence the type of skills imported to Australia in the future.

Under the points system, it is not essential for applicants to secure any or all of the available points for English to be eligible to enter Australia. For example, in May 1994 Independents required 100 points and Concessionals 95 (DIEA 1994a). Seventy of these points could be scored by possession of a recognised trade certificate/degree/diploma alone, provided that this was supported by three years professional experience. Substantial additional points could be gained by applicants being of prime workforce age. Even Independent applicants with modest English skills could secure an additional ten points for their competence in English. Similarly, Concessionals had the potential to score points for family relationship, citizenship and location.

Given this situation it is the Occupations Requiring English (ORE) — 114 occupations across a range of previously defined fields — which may in future exert powerful control over the type of skills selected within the skilled migration program. Principal applicants in non-Humanitarian categories will be ineligible to migrate, regardless of the number of points earned, if they cannot demonstrate possession of ‘vocational’ levels of English. This requirement has the potential to reduce significantly the numbers of incoming NESBs in ORE professions such as engineering, nursing, medicine and teaching, as well as within select ‘public safety’ trades (such as electricians). Simultaneously, contemporary policy may be inadvertently admitting a growing proportion of NESB applicants with recognised skills which are marginal to the Australian labour market and thus less in demand, such as philologists, architects or lawyers since they have not been defined as skilled in the ORE category. These migrants may arguably require greater English language competence to secure work, given the existence of fewer job opportunities but as these occupations are not on the ORE list, applicants will not have been required to demonstrate a high level of English in advance and many will inevitably migrate without it.

Given these two situations, DIEA may be compelled in the future to lower the English standards required of ORE candidates and/or increase the number of occupations on the list.7

**Maintenance of standards**

A fourth issue concerns the maintenance of the test development and monitoring standards established for access. The delivery of the test in its post-development stage will require careful and ongoing oversight in order to ensure that the test continues to meet accepted international standards.

**Conclusion**

The period from 1993 to 1996 has seen the progressive implementation in the skilled migration program of mandatory English language testing at overseas posts using access: as a screening tool.

Given its relatively recent introduction, little information is available to date on the consequences of the access: test, either off-shore or within Australia. In
order to evaluate its effectiveness as a migration tool, it will be necessary to undertake tracer studies aimed at investigating the extent to which it is able to predict candidates' future performance in social and vocational situations. At the same time the opinions of candidates and users of test information about its fairness and efficacy will need to be sought. In this way it should in time be possible to build up a body of data which can constitute a basis for the continuous improvement of the test.

Notes
1. This chapter is based on two articles by Lesleyanne Hawthorne previously published in *People and Place*: 'The politicisation of English: the evolution of language testing' (vol 2, no 2, 1994), and 'The politicisation of English part two: the access test and the skilled migration program' (vol 3 no 1, 1995).
2. All immigration dates refer to financial years (ie July to June), as this is the DIMA unit of planning.
3. The Australian government accepts skilled migrants in three immigration programs: the skill migration program, the family migration program, and the refugee/special humanitarian program. The skill migration program selects individuals on a points-tested basis in a range of sub-categories, the chief of which is the Independent category. The family migration program includes a points-tested, sub-category (Concessional), designed to select skilled migrants with family links to Australia (eg siblings, aunts). The refugee special humanitarian program (predominantly composed of low-skilled intakes selected on the basis of humanitarian need) also contributes a small proportion of skilled professionals to Australia, though these individuals have not been subjected to points-testing. The points test, administered by the Department of Immigration and Multicultural Affairs since the 1980s to assess the workforce potential of skilled migrants, allocates a set number of points for desired attributes. These include qualifications level, professional experience, age (with a weighting towards youth), English competence (for Independents), family relationships (for Concessions), and intended area of location. The 'pass' level fluctuates considerably over time, typically between 85 and 115 over the past decade. It is generally low for Concessional applicants. Sources: *Fact Sheet 22*, Public Affairs and Information Section, Department of Immigration and Ethnic Affairs, Canberra,
Appendix: Occupations appearing on Occupations Requiring English list

(Commonwealth of Australia Gazette, No S 33, 1 February 1993, Canberra)

Director of nursing
Orthopaedic specialist

School principal
Psychiatrist

Education manager (including
Radiologist

Dean, faculty heads of universities
Paediatrician

and CAEs)
Thoracic specialist

Commissioned police officer
Dermatologist

Engineering manager
Otorhinolaryngologist

Medical laboratory scientist
Urologist

Clinical physical scientist
Surgeon

Chemical engineer
Pathologist

Civil engineer
Specialist in rehabilitation medicine

Electrical engineer
Specialist physician

Electronics engineer
Specialist medical practitioner (not

Mechanical engineer
elsewhere classified)

Mining engineer
Dentist

Petroleum engineer
Dental specialist

Materials engineer
Hospital pharmacist

Aeronautical engineer
Industrial pharmacist

Agricultural engineer
Retail pharmacist

Naval architect
Occupational therapist

Industrial engineer
Optometrist

Engineer (not elsewhere classified)
Physiotherapist

General medical practitioner
Speech pathologist
<table>
<thead>
<tr>
<th>Anaesthetist</th>
<th>Chiropractor</th>
<th>Rehabilitation counsellor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardiologist</td>
<td>Osteopath</td>
<td>Detective</td>
</tr>
<tr>
<td>Obstetrician and gynaecologist</td>
<td>Podiatrist</td>
<td>Safety inspector</td>
</tr>
<tr>
<td>Ophthalmologist</td>
<td>Clinical psychologist</td>
<td>Electrical powerline tradesperson</td>
</tr>
<tr>
<td>Medical diagnostic radiographer</td>
<td>Educational psychologist</td>
<td>Careers counsellor and adviser</td>
</tr>
<tr>
<td>Medical therapeutic radiographer</td>
<td>Occupational psychologist</td>
<td>Counsellor (not elsewhere classified)</td>
</tr>
<tr>
<td>Nuclear medical technologist</td>
<td>Curriculum development officer</td>
<td>Journalist</td>
</tr>
<tr>
<td>Veterinarian</td>
<td>Education officer</td>
<td></td>
</tr>
<tr>
<td>Audiologist</td>
<td>Medical laboratory technical officer</td>
<td></td>
</tr>
<tr>
<td>Dietitian</td>
<td>Medical laboratory technician</td>
<td></td>
</tr>
<tr>
<td>Orthoptist</td>
<td>Dental therapist</td>
<td></td>
</tr>
<tr>
<td>Health Diagnosis and Treatment Practitioners (not elsewhere classified)</td>
<td>Medical officers and technicians (not elsewhere classified)</td>
<td></td>
</tr>
<tr>
<td>Pre-primary school teacher</td>
<td>Aircraft pilot</td>
<td></td>
</tr>
<tr>
<td>Senior teacher</td>
<td>Air traffic controller</td>
<td></td>
</tr>
<tr>
<td>Infant-primary teacher</td>
<td>Flight service officer</td>
<td></td>
</tr>
<tr>
<td>Primary school teacher</td>
<td>Flight engineer</td>
<td></td>
</tr>
<tr>
<td>Senior teacher, secondary school</td>
<td>Airways surveyor</td>
<td></td>
</tr>
<tr>
<td>Secondary school teacher</td>
<td>Airworthiness surveyor</td>
<td></td>
</tr>
<tr>
<td>Resource teacher</td>
<td>Examiner of flightcrew</td>
<td></td>
</tr>
<tr>
<td>Teacher of disabled or gifted children</td>
<td>Aircraft navigator</td>
<td></td>
</tr>
<tr>
<td>Special education teachers (not elsewhere classified)</td>
<td>Trainee air traffic controller</td>
<td></td>
</tr>
<tr>
<td>University or CAE lecturer</td>
<td>Registered general nurse</td>
<td></td>
</tr>
<tr>
<td>University or CAE tutor</td>
<td>Registered community health nurse</td>
<td></td>
</tr>
<tr>
<td>Head of school (TAFE)</td>
<td>Registered midwife</td>
<td></td>
</tr>
<tr>
<td>TAFE teacher (trades)</td>
<td>Registered mental retardation nurse</td>
<td></td>
</tr>
<tr>
<td>TAFE teacher (general education)</td>
<td>Registered psychiatric nurse</td>
<td></td>
</tr>
<tr>
<td>Flying instructor</td>
<td>Student registered nurse</td>
<td></td>
</tr>
<tr>
<td>Ground school instructor</td>
<td>Registered nurse (not elsewhere classified)</td>
<td></td>
</tr>
<tr>
<td>Nurse educator</td>
<td>Police supervisor</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Policeman/woman</td>
<td></td>
</tr>
</tbody>
</table>
Introduction

This chapter provides a brief history of the development of the access: test. It begins with an outline of the stages involved in the test development — from the formation of a consortium to the first administration of the access: test — and an overview of its structure. Some of the challenges which had to be addressed at the initial design stage, including the potential conflict between validity and practicality are also discussed. The process by which each of the four test modules (Listening, Oral Interaction, Reading and Writing) was developed is then described in greater detail and the essential features of each are summarised. Some of the issues and dilemmas which arose in task selection, item design, scoring and analysis are considered, as are some of the key practical and theoretical influences which helped to shape the final form of each test.

Background

In September 1992, the Minister for Immigration and Ethnic Affairs at the time, Mr Gerry Hand, announced that NCELTR had been commissioned to develop a new test in order to provide Australian migration officers overseas with an independent and objective assessment of applicants’ English language proficiency. The testing instrument had to serve three different purposes depending on the type of candidate (See Chapter 1). The test had to determine:

- whether candidates in the Occupations Requiring English category met the mandatory level of vocational proficiency in English;
- the level of language skills for other candidates to determine the points they would be awarded under the Numerical Migrant Assessment System (NUMAS);
In 1995, ongoing responsibility for the construction and delivery of the test was transferred to a consortium made up of the International Development Program of Australian Colleges and Universities (IDP) in partnership with LTACC at Griffith University. This chapter therefore describes the development and administration of access: only up to the end of 1994.

Preliminary stages

The original access: contract called for the delivery of the test in nine locations by April 1993. Preliminary specifications were drafted in August 1992, and item development began immediately. The basic components of the specifications were decided upon by the TDC to ensure consistency across all test modules. Early drafts of specifications for each module were circulated for review, and final drafts were ratified by the TDC. At this stage, given the urgent need to produce items quickly, the specifications were aimed primarily at item writers, rather than test users. However, the test developers were also aware of the need to produce a ‘user-friendly’ description of the aims and content of the test had to be made available to test as soon as possible.

Specifications

Because of the very short time frame for test development, it was intended that the specifications, rather than constituting ‘a blueprint to be followed mechanically’ (Pollitt 1993: 1), would be reviewed and revised throughout the first year on the basis of feedback from the development and trialling process. This is consistent with the current view that specifications should be subject to re-evaluation and modification once they have been operationalised and tested (Pollitt, op. cit.).

The access: specifications aimed to describe the following dimensions for each of the four modules:

- purpose of the test;
- target population;
- topics (including cultural content);
- levels;
- test structure;
- language operations/skills tested;

NCELTR also had responsibility for the development of the access: listening module while the development of the reading and writing modules was undertaken by the NSW Adult Migrant English Service (AMES). The National Languages and Literacy Institute of Australia (NLLIA) Language Testing Research Centre (LTRC) at the University of Melbourne — assisted by staff from the Royal Melbourne Institute of Technology (RMIT) Centre for English Language Learning (CELL) — was responsible for the construction of the oral interaction module. The NLLIA Language Testing and Curriculum Centre (LTACC) from Griffith University provided advice and assistance with the development of test specifications and items. To facilitate the setting of performance standards, LTACC also undertook an investigation of the relationship between the scores of candidates who had been rated at the key ‘functional’ and ‘vocational’ levels on the Australian Second Language Proficiency Rating (ASLPR) scales and their scores on pilot access: tests.

The test development process was overseen by a Test Development Committee (TDC) chaired by NCELTR and consisting of representatives from all participating institutions. Separate editorial committees consisting of members of the TDC along with other colleagues with expertise in the particular language skill concerned were formed to oversee the development of each of the test modules. Five versions of the test were developed by the NCELTR-led consortium between September 1992 and December 1994.
occupational and cultural backgrounds and hence not easy to define precisely. This made it necessary to find an onshore trial population as close as possible to the profile of candidates offshore. Initial trialling of the oral interaction was therefore conducted at the Royal Melbourne Institute of Technology (RMIT) on a sample of ninety-four students consisting of recently arrived migrants and professionally-qualified overseas students. These students represented the range of proficiencies which were included in the access: test, although most were clustered around the key vocational level (see Chapter 5 for further details of the candidate profiles).

In February 1993, all modules of the access: test were piloted with a population similar to that at RMIT, but with a wider band of proficiency levels — this time ranging across both functional and vocational levels. The test modules were then modified in the light of the analysis of the results from the trialling, together with comments by raters and test takers before the first delivery in April 1993. New versions were trialled in a similar manner and on similar populations as they were developed.

The trialling process included a series of feedback questionnaires completed by the test candidates for all modules, the raters of the oral interaction and writing modules, and the oral interaction interviewers. The purpose of these questionnaires was to elicit respondents' opinions on the appropriateness and fairness of the tests, and included questions on the comparability of the two formats of the oral interaction. The results of the questionnaires provided feedback on certain aspects of test validity, specifically face, content and construct validity. Feedback also allowed revisions to tasks in terms of clarification of rubrics and time allowed for response to be carried out, thus improving test reliability.

Describing proficiency levels

The results of the test are reported in the form of a Proficiency Profile which describes candidates’ ability in each of the four skills on a six-point scale. In order to assist users in interpreting the skill levels, the Proficiency Profile provides a set of simple descriptors which summarise in general terms the types of performances which could be expected from candidates at different levels (see Table 1).

In order to assist migration officers in making decisions, the skill levels for each of the four skills are added together. The total thus obtained is known as the Assessment Score. In order to allocate points, this score is then related...
back to the descriptions of language proficiency (labelled Vocational, Functional, etc) which are defined in the Immigration Regulations.

**Table 1: Proficiency profile for access: six-point scale**

<table>
<thead>
<tr>
<th>Level</th>
<th>Read</th>
<th>Write</th>
<th>Understand</th>
<th>Speak</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level Six</td>
<td>You can read and understand a wide range of English texts easily and with good comprehension; you can write English appropriately and with quite a high degree of accuracy for a range of purposes; you can easily understand spoken English in a wide variety of situations; you can speak English appropriately and with quite a high degree of accuracy and fluency in most contexts.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level Five</td>
<td>You can read and understand a variety of texts in English with reasonably good comprehension; you can write English well enough to communicate effectively for most purposes; you can understand spoken English quite competently in a range of situations; you can speak English fairly fluently and accurately in a range of contexts.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level Four</td>
<td>You can read and understand English texts about familiar topics; you can write English well enough to communicate ideas or information for a variety of purposes but you make some errors; you can understand spoken English about familiar topics; you can speak English fairly fluently and accurately in a range of contexts.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level Three</td>
<td>You can read and extract basic information from everyday written texts in English; you can write enough English to communicate simple messages but with frequent inaccuracies; you can understand enough spoken English to comprehend some of the main points in simple conversations about familiar topics; you can speak English well enough to handle everyday communication adequately, despite some errors.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level Two</td>
<td>You can read and understand some words and phrases in very simple everyday texts in English; you can write enough English words and phrases to communicate a restricted range of very simple information on familiar topics but with many inaccuracies; you can understand a limited range of common English words and phrases in simple conversations; you can speak enough English to have a very elementary conversation but with many errors and hesitations.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level One</td>
<td>This means that you cannot read, write, understand or say anything in English or that you know only a few common words and phrases.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Considerations in test construction: the listening module

Authenticity and validity

The large-scale testing of listening comprehension is fraught with problems of validity. In the first place, a good deal of listening in everyday life (e.g. listening to radio or television programs) does not require a specific response on the part of the listener. Yet in a formal listening test, the test developer is forced to elicit some kind of productive performance in order to establish the extent to which understanding has occurred. Not only is this mode of testing inauthentic in so far as it does not correspond to natural language use, but it also heightens the risk of the confounding of skills since many item formats demand a spoken or written response, usually in the test language rather than the candidate’s first language.

A further threat to validity, which arises from the practical constraints of group testing, is that standardised listening tests tend to focus on non-interactive listening tasks. Candidates typically listen to pre-recorded texts and tick boxes, circle alternatives or write short answers on the test paper. A good deal of real-life listening, however, happens in the context of oral interaction where listening and speaking ability are closely interconnected; that is, a person cannot carry on a conversation effectively without understanding what the interlocutor is saying. This type of conversational listening ability is rarely sampled in language tests, though some speaking tests may include listening ability as one of the rating criteria in an oral proficiency rating scale.

An additional factor affecting test validity is the extent to which it is possible to provide adequate contextualisation for listening texts. Nowadays a good deal of emphasis is given to the importance of providing adequate visual context to learners when teaching listening comprehension so they are able to activate their content schemata (videotaped materials, for example, are routinely used in language teaching classrooms for this purpose). In the interests of validity and authenticity it would have been desirable to use video-recorded aural texts in the access: test rather than audio recordings. However, this option was not feasible owing to the unavailability in many test sites of the required equipment and the lack of suitable venues where video-based testing could be carried out.

Although the test developers were aware of all of these problems concerning the ability to obtain an adequate or authentic sample of listening ability from a non-interactive test, the logistics of test administration made it necessary to develop the test in a form which could be administered via audio cassette to large groups. The main focus of the test, therefore, was on non-interactive listening skills and in particular on candidates’ ability to extract information from a range of sources which they might be expected to encounter in everyday life in Australia. However, in an attempt to incorporate an element of conversational listening in the test, the TDC decided to include in each version of the oral interaction module at least one task containing an extended passage of spoken input which candidates had to understand in order to produce an appropriate response.

Skills

Although item writers are likely to use a listening passage rather than a list of skills as a point of departure for item development, a taxonomy of listening skills adapted from Rost (1990) and Weir (1993) was included in the specifications given to them (see Chapter 3 where the taxonomy is reproduced). It was hoped that such a taxonomy would enable item writers to check retrospectively (albeit in an impressionistic way) the extent to which various skills were being sampled by particular listening tasks. It was also anticipated that the taxonomy could serve as a basis for any content analysis that would later be carried out as part of research into test validity.

However, it should be noted that the relationship between listening skills and task performance is the subject of ongoing debate in language testing. Moreover, a number of research studies have cast doubt on the extent to which it is possible to match listening test items to specific skills (Buck 1990, 1991), given that listeners may deploy a range of skills simultaneously to extract meaning from an aural text. The implications of these findings for the access: test are discussed in Chapter 3.

Test structure and grading

The listening test was constructed in five sections beginning with short, relatively simple passages and progressing to longer and more difficult texts. The question of what makes a listening text ‘difficult’ is, of course, a complex and controversial one and relates to the interaction of a wide range of learner, input and task factors (see Chapter 3). Broadly speaking, however, the following variables, which have been identified as key factors affecting task difficulty (Brindley and Nunan 1992), were taken into account in attempting to grade the passages (a
good deal of further research remains to be done, however, before the specific effects of each of these variables can be gauged):

- speed of delivery;
- length of text;
- density of information in text;
- amount of repetition and redundancy in text;
- amount of slang/idiom used in text.

As the listening test was to be approximately thirty minutes long and candidates were to be given only one hearing of each text, it was imperative that each task was adequately contextualised through the use of visuals/sound effects and a spoken and written orientation to the task. Repetition and recycling were also used to make the earlier passages more accessible to lower level candidates. Rubrics which trialling revealed to be ambiguous or difficult to understand were revised before being included in the live version of the test. Other aids to accessibility which were built into the test included:

- Where a new item type is introduced, an example was given.
- A short pause is given after each section to allow candidates to check their work, and a longer pause is left at the end of the test, before candidates are told that they have reached the end.

**Text selection**

Most of the spoken texts used in the access: listening module are representative of varieties of general Australian English, although some speakers have slight ‘foreign accents’. Some passages are based on authentic recordings, while others were scripted specifically for the test. Here again, the developers were forced to sacrifice a degree of authenticity in the interests of practicality by carrying out some degree of editing on most of the authentic listening material recorded. While it is clearly desirable to use ‘real life’ recordings which reflect the target language use situation, it was found that unedited authentic texts were frequently unsuitable for use as testing material. Sometimes they were too long; at other times they contained either too much or too little information, or would have required extensive contextualisation. Even some texts which were capable of generating a considerable number of items had to be edited when it was found, for example, that the items followed one another too closely and thus placed an overly-demanding processing burden on candidates or that a text contained specific cultural references which would be difficult for some candidates to interpret.

The decision to include non-standard accents provoked some debate within the TDC, since some members felt that accented English could disadvantage candidates unfamiliar with the accent in question or privilege those from that particular language background. However, in the light of the findings of Ross and Langille (see Chapter 4), there is some justification for this decision.

The kinds of listening texts which were used in the test involve the use of language for both transactional purposes (conveying factual or propositional information) and interactional purposes (expressing social relationships and personal attitudes). Text types include conversations, announcements, service encounters, answering machine messages, directions, lectures, narratives, personal reports, advertisements and talkback exchanges.

**Item formats**

In order to minimise the well-documented effects of particular item types on test performance (Bachman 1990), the developers of the listening test tried to ensure that a range of task/item formats was represented in each version. The following is a description of the final structure.

Each of the five sections of the test were designed to contain between six and ten items. Section 1 was constructed in two parts, the first part being a dialogue or a conversational exchange, using item formats such as form filling or completing a chart testing candidates’ ability to extract relatively straightforward specific information relating to such things as times, numbers, names, places, objects. The second part of Section 1 required the candidate to follow directions/instructions or to extract specific information in relation to a visual stimulus. This item format minimises the need for written responses and lends itself to activities such as ticking boxes or matching pictures.

Section 2 incorporated a longer text (up to three minutes), usually consisting of an interview, conversation or monologue aimed at testing candidates’ ability to identify the main points in a text using formats such as gap filling and short answers.
Section 3 included a conversation or dialogue of at least three minutes consisting of an exchange of information between two people. This section is also aimed at assessing candidates’ ability to extract specific information from a text employing item formats such as grid completion and short answer questions. However, it was desired to be longer and to test more complex information than Section 1 (for example, candidates might be asked to specify a reason for an action rather than merely identify a number or name as in Section 1).

Section 4 centred around a monologue, talk or interview, up to four minutes long on a non-specialised subject (e.g., the environment, immigration, health, education). Skills targeted included the ability to understand the gist and main details of a text when these are explicitly stated. Item formats ranged from short answer questions requiring synthesis of information to gapped summaries.

Section 5 consisted of a longer conversational dialogue (up to five minutes) spoken at a faster rate of utterance, usually involving two or more participants expressing different points of view or attempting to negotiate a transaction or interaction. It was intended to test candidates’ ability to understand meaning which is not explicitly stated in the text, in particular to identify the speaker’s point of view, line of argument or attitudes when these are not expressed directly. Following trialling, it was decided to use some multiple-choice items in this section, despite their well-documented shortcomings. The results of item analysis showed that these types of items demonstrated acceptable psychometric properties and they subsequently proved to be good discriminators at the upper levels of proficiency.

**Test analysis**

Test scores from trialling were analysed using the Rasch-based QUEST program (Adams and Khoo 1992) which provides a range of item level statistics including facility values, point-biserial correlations and a measurement error for individual items. An overall test reliability index is also available in the form of a ‘reliability of case estimate’ statistic — the Rasch equivalent of the Kuder-Richardson 20 formula for internal consistency. The QUEST program output also includes a ‘fitmap’ showing the extent to which each item fits the response patterns predicted by the Rasch model along with an item/ability map which displays item difficulty levels and person ability estimates on the same scale.

On the basis of the information provided by the QUEST program following trialling, misfitting items were identified. The listening test subcommittee discussed possible reasons for the misfits and problematic items were modified or sometimes dropped completely. The elimination of items in a listening test, however, can create logistical problems since a deletion may sometimes result in a lengthy pause in the recorded text. This may be disconcerting to candidates if items have been more or less evenly spaced throughout the test up to that point. Conversely, the creation of a new item which follows closely on an existing one may mean that there is insufficient time for candidates to process the necessary information and respond to the question. This, too, can present problems for candidates, who may become disoriented if they ‘lose their place’ in the text and thus fail to process items which follow. Care was therefore taken to ensure that adequate processing time was allowed between items.

**Scoring procedures**

Although there are some multiple choice items, the majority of items in the access listening module are open-ended, short answer questions which are then clerically marked. A detailed scoring key containing a list of acceptable responses based on extensive trialling is provided to markers. However, even after trialling, it was not always possible to produce an exhaustive list of plausible responses and as the papers of each new cohort were marked, further plausible answers would come to light. Markers were asked to identify these and to bring them to the attention of the Chief Examiner who adjudicated on their acceptability, in consultation with other members of the listening subcommittee. In this way, a comprehensive set of marking guidelines was gradually developed. This experience nevertheless highlights the inherent difficulties in deciding on what is a ‘plausible interpretation of a text’ and strengthens the case for using procedures for item development such as those suggested by Weir (1993: 111) who recommends that items should be based only on those main points and details identified by native and non-native listeners through mind-mapping and note-taking.

Not surprisingly the issue of acceptable responses also arose during the development of the reading test, and this is discussed later in this chapter.

**Summary**

As Weir (1993: 100) points out, ‘there is a tension between the describable and the testable in the assessment of listening’. Current views of listening ability suggest that it is a highly complex and interactive process which involves the use of a wide variety of verbal and non-verbal cues in interpreting messages.
It is not easy, however, to tap this ability in the space of a short test which has to be delivered via a disembodied tape recording. As far as possible the developers of the access: listening module, in the selection of content and in the use of a variety of item formats, attempted to broadly sample the kinds of listening tasks that candidates might have to undertake in everyday life in Australia. As already pointed out, it would have been more consistent with current views of the listening process to use texts which were better contextualised through the use of video technology. Unfortunately, because of the tight development schedule and the practical constraints of group testing in often under-resourced locations, this was not possible and the test that resulted turned out to be a somewhat more traditional test than the developers had originally hoped to produce.

Considerations in test construction: the oral interaction module

Interview or tape-based? Advantages and disadvantages

Constraints upon test delivery and administration were a major factor influencing the decision to develop two forms of the oral interaction module — an interview-based format and a tape-based format. This decision had considerable consequences for the subsequent design and structure of the test, since the two formats needed to be equivalent, and ongoing research into the comparability of the two tests was carried out during the two years in which access: was developed at the LTRC at the University of Melbourne. A detailed discussion of this research can be found in Chapters 5–7 and in O’Loughlin (1995; 1996).

The two different formats for testing oral interaction have different advantages and disadvantages. The live interview allows for more interaction and can therefore contain more interactive tasks. It can also be rated at the time of the interview by a trained interviewer/rater (although this was not an option which was adopted for this test). However, interviewers are variable in the way they perform, and this variability can be compounded when the interviewer is also the rater, thus introducing potential measurement error into candidate scores. In addition, the interview format is expensive as each candidate must be interviewed by a single interviewer.

The tape-based format, however, is limited in interactional terms because of the absence of a live interlocutor, and all tasks must be one-way. On the other hand, this format has the advantage of being cheaper because many candidates can do the test at the same time. It also ensures that all candidates receive the same input. It is quicker to rate because all sections begin at exactly the same point on the tape and raters can identify these rapidly. However, it does not have the same face validity as the interview format which more closely resembles natural communication.

Whether a candidate took the tape-based or live test had to be determined partially by logistical considerations such as the capacity of the test centre (Hong Kong, for example, typically had large numbers of candidates) and the facilities available (Hong Kong was able to provide language laboratories suitable for large scale group testing). Availability of trained native speaker interviewers was another key factor in deciding which format to administer since it is much more difficult to deliver the live test in locations where native speakers are hard to find.

Task selection

The design of the oral interaction module involved a number of challenges. At the most fundamental level there were issues such as the number of tasks which should be included and the range of language functions/forms which should be elicited by these tasks. There was also the question of how to make the tasks as realistic as possible — as already noted, authenticity is not always easy to achieve in the test situation. In an attempt to achieve as much authenticity as possible, test developers endeavoured to include a wide variety of tasks which would reflect the variety of situations in which adult Australians might find themselves in the everyday course of their work, social and educational pursuits. The specifications focused, in particular, on the kinds of language which might be expected to occur in vocational settings.

In order to ensure that candidates were not disadvantaged by encountering tasks which were unfamiliar, or which they found particularly difficult for one reason or another, the test contained a range of tasks which gave all candidates the opportunity to use their language resources to their full potential. In deciding on the number of tasks, however, a difficult balance had to be maintained. On the one hand, it was necessary to ensure that there were not too many tasks, so that candidates became tired (which could mean they were disadvantaged towards the end of the test). At the same time, there had to be enough tasks to elicit a rateable sample of language from the candidates.
Task types

While a number of different task types were identified in the specifications, the limits were imposed by the need for similar task types to be included in both the tape-based and live interview formats of the test in order to allow for ongoing analysis of comparability. It was also for reasons of comparability that a decision was made to use mainly one-way exchanges which were largely monologic in nature. Thus candidates would be required to communicate information in response to a prompt. However, since the interview format of the test allows for more authentic interactive tasks, a role-play was included in the interview format of the test. The parallel task on the tape-based format was a response to an answering machine message. The format of the two tests as they were originally designed is set out in Table 2.

Table 2: Format of access: oral interaction module

<table>
<thead>
<tr>
<th>Section</th>
<th>Live interview</th>
<th>Tape-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Section 1</td>
<td>warm up</td>
<td>warm up</td>
</tr>
<tr>
<td>Section 2a</td>
<td>picture description/comparison</td>
<td>picture description/comparison</td>
</tr>
<tr>
<td>Section 2b</td>
<td>narrative</td>
<td>narrative</td>
</tr>
<tr>
<td>Section 2c</td>
<td>(not applicable)</td>
<td>summary</td>
</tr>
<tr>
<td>Section 3</td>
<td>role-play</td>
<td>telephone answering machine</td>
</tr>
<tr>
<td>Section 4</td>
<td>explication/process task</td>
<td>explication/process task</td>
</tr>
<tr>
<td>Section 5</td>
<td>interview(dialogic)</td>
<td>interview (monologic)</td>
</tr>
</tbody>
</table>

The warm-up section of the listening module consisted of a few personal questions designed to relax the candidate. In Section 2a, the candidate was asked to look at a picture of a familiar scenario and to describe the scene. The candidate was then asked to look at another picture and to compare it with the first, and was then asked to give personal opinions and/or preferences. The narrative task in Section 2b consisted of a series of eight pictures which told a story that the candidate was required to relate. Section 2c, which occurred only in the tape-based format, asked the candidates to listen to a conversation and then to summarise it. Section 3 consisted of a role-play in the live interview, and a response to a telephone answering machine message in the tape-based format. Section 4 required the candidate to either describe a process, or to discuss some data summarised in the form of a chart or graph. In the final section the candidate was asked a series of vocationally and/or culturally oriented questions.

Tasks for both formats of the test were designed in tandem to ensure the greatest possible comparability. For each version of the test, three or four of each type of task were developed and these were presented to an editorial committee for comment. The committee consisted of two members of the TDC, a member of the LTRC who was not involved with access: on a daily basis, and two external members with expertise in both ESL and immigration issues. The committee scrutinised all tasks. Unless rejected outright, tasks were revised in line with comments from the committee and then subjected to pre-trialling on a small number of non-native speakers to ensure that picture content, rubrics and questions were clear and elicited the expected response type. Tasks which survived this process were then incorporated into one of the two formats. Once both formats of a version of the access: test were complete, they were trialled along with the listening, reading and writing modules (see below for further discussion of trialling).

Training and selection of interviewers

While the test was delivered in the tape-based format in Hong Kong, where substantial numbers were taking the access: test, in other centres the interview format was used. It was therefore necessary to identify ways in which interviewers could administer the test to ensure maximum fairness for all candidates. One way of doing this was to structure the test in such a way that it reduced interviewer variability. This is an important consideration in light of several recent investigations of interviewer behaviour which have identified the interviewer as a significant variable in test administrations (Ross 1992; Ross and Berwick 1992; Lazaraton 1996; Young 1995; Young and Milanovic 1992).

Careful selection and training of interviewers is also essential for equity. Since interviewers were chosen by centre coordinators overseas, the access: team set down clear guidelines concerning interviewer competency: Interviewers were required to be native speakers, with formal training in the teaching of English as a Second Language, and with at least two years experience teaching at a range of proficiency levels. A training video and booklet were developed and sent to all overseas centres to be used in the training of interviewers. Training sessions were held by the coordinator of each overseas centre prior to the administration of the test to ensure that the interviewers were familiar with the contents of the test. The training session involved interviewers watching and discussing the video which depicted a series of sample test performances, and familiarisation with the version of the test to be administered. For security reasons, interviewers...
Raters were selected on the basis of their qualifications and experience. Like the interviewers, they were required to have formal qualifications for teaching English as a Second Language, and to have at least two years teaching experience which covered a range of levels of proficiency.

Their extensive training program consisted of two workshop sessions in which audiotapes of ‘benchmark’ performances (considered to be typical of each level) were played to the raters. These benchmark performances were identified on the basis of ratings given by four or five experienced oral test developers. Raters scored these individually and the ratings were then discussed in a plenary and/or small group session.

After participating in the training workshops, each rater was required to assess a set of thirty tapes — fifteen interview format and fifteen tape-based. These were then analysed to determine whether the rater was rating consistently. The statistical program which was used to analyse the results, FACETS, adjusts for rater harshness or leniency. Therefore, the crucial question in rater training is not so much whether raters agree with each other (inter-rater reliability) but whether each rater is consistent in his/her own ratings (intra-rater reliability). Subsequent to training, intra-rater consistency was monitored following each administration through the use of bias analysis, a technique which enables the identification of tasks or ratings which deviate from expected response patterns.

**Test analysis**

The FACETS program used to analyse the test results is based on the Rasch model (see notes) but extends the model to take rater characteristics into account. The ability estimate for each candidate — expressed in units called logits — is produced through an analysis of the interaction between facets, that is aspects of the test setting which may influence the result the candidate achieves on a test. For the purposes of the access: oral interaction module, the candidate, the rater, and the item were all identified as facets, each contributing to the candidate ability estimate. In addition to the ability estimates, the analysis provides a logit value for each facet identified, such as the harshness of the rater or difficulty of items. This enables the relative difficulty of items or the relative severity of raters to be described and compared.

A technique known as bias analysis may be used to investigate whether particular raters rate more or less harshly or leniently on a particular task or criterion, such as fluency or grammatical accuracy. In cases where two formats of a test
are being used (as in the interview and tape-based formats of access) such an analysis can provide an overview of each rater’s characteristics with respect to the different formats of the test. Bias analysis is, therefore, a useful tool both for monitoring tasks and improving rater consistency, thus reducing the proportion of error variance in candidates’ scores.

Summary

The assessment of oral interaction skills presents an ongoing challenge to test developers. Issues such as authenticity of both the stimulus material and the nature of the interaction elicited are in need of a good deal of further research. Interviewer variability in the test situation is another important issue which has only recently begun to be seriously addressed. In addition, the subjective nature of the assessment process demands constant and vigilant monitoring of rater behaviour and consistency. For access: there were the additional challenges of monitoring interviewers at overseas posts from a distance, and the necessity of producing two comparable formats. Although this constraint resulted in a test which may have lacked some degree of interactivity and spontaneous discussion, the need to be fair to all candidates, regardless of country of origin, was deemed to be of paramount importance.

Considerations in test construction:
the reading module

Text/task selection

Reading tests, rather more than speaking or writing tests, offer an opportunity to include an extremely varied array of text types, although not all text types will generate useable items or discriminate usefully between candidates. At the same time, however, a test which is being developed in multiple versions must have a consistent internal structure. A major challenge in the development of the access: test was therefore to ensure consistency across different test versions in terms of the kinds of texts which were represented, without unduly constraining item writers in their text selection.

The specifications for the reading test initially identified five categories of types of texts. These were then reduced to the following four categories of functional purpose and each version of the reading module was constructed with one text from each category:

- a general human interest category including texts which provide non-technical information intended to be of interest to a general audience, for example, tourist brochures, pamphlets or advertisements;
- texts which provide news, information or instructions for a targeted audience, for example school newsletters, office memos, circulars, information or instructions for a product;
- texts which provide semi-technical or semi-scientific information, for example, short reports, newspaper or magazine articles;
- texts which contribute to public debate or discussion around a particular issue, for example, letters to the editor, newspaper articles or flyers from interest groups.

Choice of text was also constrained by specifications of length. The reading texts were expected to be approximately one page in length (between 400 and 700 words depending on the size of typeface), but allowing for some leeway to maintain the authenticity of texts. Where a shorter text was included in one section, there was an expectation that this would be balanced by a slightly longer text in another. There was also the option to include a series of two or more short texts of the same kind as a compilation text (e.g. a number of advertisements).

It is clearly essential in reading tests that the text is pitched at an appropriate level of difficulty for the candidature. This means that a number of factors need to be taken into account, such as the topic, the length, the explicitness of the discourse structuring, the relative frequency or technicality of lexical items, the amount of repetition and redundancy, the degree of abstraction in the grammar as well as visual factors such as the layout and design of the text and accompanying visual elements.

A suitable text must also generate sufficient items of particular kinds, including items across a range of difficulty levels. Information brochures, for example, while providing attractive and interesting content, often proved difficult texts for generating items of sufficient difficulty. Such texts were designed, after all, to present factual information in a straightforward and highly accessible manner and therefore did not tend to generate items which required candidates to go beyond the text.
Skills

The constraints of the test-taking conditions and the need to create reliable items necessarily narrowed the construct of reading that could be tested. It was not possible, for example, to test for particular reading strategies or to accommodate personal responses to texts. In order to encourage item writers to sample a range of language operations and to design tasks which varied in difficulty level, a list of language operations was included in the specifications. However, as already noted in relation to the listening test, it cannot be assumed that there is an unambiguous relationship between the items and the skills which the item writers consider they are testing. Findings from a number of research studies of reading tests (such as Alderson and Lukmani 1989; Alderson 1990) suggest that there is often little agreement amongst test developers or teachers as to which operations are required for which items. Nor is there agreement on a hierarchy of difficulty in language operations. Think-aloud protocols of test takers (Anderson et al 1991; Wiigh 1996) have also indicated that different candidates use different strategies and language operations to arrive at the same answer (and likewise will arrive at different answers via the same path).

Topics

It was decided at an early stage in the test development that the content of the reading test would be based on natural language use in the target language situation (with modifications as required). The specifications stated that ‘the contexts would reflect those encountered by adults in Australia as they interact in English in social or community settings or in education, vocational training or employment’. However, the potential diversity of candidates’ cultural, educational and employment backgrounds presented some challenges when it came to choosing appropriate topics. On the one hand, test items could not assume prior knowledge of a particular field of work or study, nor could they assume familiarity with Australian society and culture. This meant that as far as possible test developers had to try to construct items where the information and ideas required were retrievable directly or by inference from the text itself.

A further constraint on the choice of topic, and one that was reasonable given the context and purpose of the test, was that texts should not present an unnecessarily negative image of Australia. It was also obviously important to avoid topics which could be personally distressing or culturally offensive to candidates. For this reason, texts on topics such as social violence and crime (or even ‘natural’ violence such as shark attacks!) were generally not considered appropriate for inclusion in the test.

Despite these constraints, it was also important to try to ensure that the test sampled a range of different topic areas and text types. In this context Hughes (1989: 51) advises:

for content validity and for beneficial washback, the important thing is to choose widely from the whole area of content (represented in the specifications). One should not concentrate on those elements known to be easy to test. Succeeding versions of the test should also sample widely and unpredictably.

Item development and pretesting

Given the need to very carefully consider texts in the light of the criteria listed above, mechanisms were set up to vet texts before the time-consuming process of developing items began. Test item writers were asked to submit potential texts to the reading test development coordinators for review and approval before proceeding with item writing. The texts submitted were accompanied by a commentary including a brief summary of the functional purpose of the text and the specific topic(s), notes on the probable item types that would be used and a few sample items. On the basis of this information, the coordinators could then decide whether to recommend further item development.

Once the text had been approved, the item writing process began. Item writers were required to submit more items than would finally be required so that any that were misfitting or rejected on other grounds could be omitted. Completed test sections (texts and items sets) were initially reviewed by the reading test coordinators and revisions were made. The test tasks were then referred for moderation and/or formal trialling. The moderation process involved a group of experienced item writers who did the test and closely scrutinised all items. Following this exercise, discussions focused on clarity of rubrics, text difficulty, problems with item design or answer keys. If additional plausible responses to short answer questions were identified these were added to the answer key.

As a result of feedback from moderation and trialling, various kinds of changes were made to the texts themselves. These included the replacement of low frequency lexis with more high frequency terms; the editing out of colloquial expressions or the insertion of some explicit cohesion to assist the reader to track
the discourse structure of the text (for example, to follow the logical connections being established). The names of real people and places also had to be changed.

Newspaper articles, especially news stories, presented many difficulties for constructing usable items. As with the texts used in the listening test, they were often too long and required a good deal of editing. Some texts needed to be re-ordered to make them more coherent. Once shortened and edited, however, longer articles proved to be a useful source for the ‘debate and discussion’ category of text.

Prior to trialling items were carefully scrutinised by the TDC in the light of guidelines for item evaluation of the type proposed by Weir (1993) and feedback was provided to item writers. This feedback ranged from outright rejection of items (eg when an item could be answered without reference to the text), to suggestions for editing (eg the modification of multiple choice distracters to ensure that they were of equivalent length).

Test analysis

Test scores from trialling were analysed using the QUEST program (see the section ‘Test analysis’ in ‘Considerations in test construction: the listening module’ earlier in this chapter). Decisions on whether to retain, modify or delete items were made using the information provided by the QUEST output in conjunction with feedback from the reading test editorial subcommittee, the TDC and the trial candidature. The item fit map enabled the developers to check that the range of items used was commensurate with the abilities of the candidature and that an appropriate range of item difficulties was represented in the test.

On the basis of the information generated by QUEST, an attempt was made to ensure that each text began and ended with easier items so that candidates were introduced to the task and finished on a positive note. In practice, however, this aim was difficult to achieve as items had to follow as far as possible the sequence of information in the text and were thus not easily reordered according to their difficulty level.

Item formats

Among the item formats included in the reading test were short answer questions (SAQs), multiple choice questions (MCQs), true/false/information not given (T/F/NG); summary cloze; matching (eg opinions to people, summary statements to paragraphs, vocabulary items to synonyms) and completing a grid of information from a text.

Each of these item types presented particular difficulties and constraints. SAQs (limited to no more than five words) were favoured by item writers as they were considered fairly easy to construct. However, such items pose particular problems in that they require an explicit answer key listing the full range of acceptable responses in order for the test papers to be clerically marked. As in the case of the Listening module, these lists rarely proved to be exhaustive because further plausible responses tended to emerge during the live marking. This necessitated the presence of a member of the test development team to decide whether or not a response was acceptable and should be added to the list.

MCQs were also popular with item writers despite cautions given in training sessions. Weir (1993: 82), for example, expresses ‘severe reservations … about employing a multiple-choice format in the testing of comprehension, either spoken or written’. Despite a belief on the part of item writers that this item type was easy to construct, many of the MCQs submitted were rejected in moderation processes or subjected to extensive editing and revisions. The main problems encountered were in constructing distractors which represented plausible interpretations of the text and were of equivalent length and complexity. It was also necessary to pay close attention to the language used in the stimulus so that the item did not become a more challenging reading task than the text. Here the item statistics from trialling were a key element in determining whether or not items needed revision.

The summary cloze item type was less frequently tried, although some items of this type survived the moderation and trialling process. The relative paucity of such items can be explained by the fact that they proved extremely difficult to construct in such a way that the candidate could not either complete the summary cloze without reading the original text, or simply find and copy items from the text. It also proved difficult to construct such items without requiring the candidate to read considerably more text than in the original.

Item formats requiring matching of textual elements were effective with certain text types. Discussion texts, for example, were often accompanied by item types which required candidates to match opinions to people mentioned in the text. However, these items tended not to discriminate well at higher levels. Items which called for candidates to match vocabulary items and synonyms to infor-
mation and ideas in the text did not produce many useable items since such items could frequently be answered on the basis of vocabulary knowledge alone.

The ‘true/false/information-not-given’ items also proved problematic in some ways. The third category of ‘information not given’ was necessary to ensure that candidates could not simply guess with a fifty per cent chance of obtaining the correct response. However, this introduced the problem of deciding whether information might be arrived at by inference from the text, was in fact ‘not given’ or was implicit in the text and thus true. Feedback from trial candidates also indicated that this item type was potentially confusing.

Although each version of the test had to contain a minimum number of items, item writers were advised to resist the temptation to create items which were peripheral to the significant points of the text, though they may have been quite straightforward to design. The number of item types used was to be restricted to no more than two per text.

Rubrics

The rubrics for each task indicated an approximate amount of time that candidates were advised to spend on that task. Each task began with a statement of ‘background’ which situated the text in a particular context. The rubric generally directed the candidate to look first to the questions and then to read the text to answer them. While the developers understood that different candidates would choose to approach the task in different ways, it was thought that it would assist candidates if they were given the opportunity to at least glance at the item types before proceeding to read the text. Rubrics were edited in order to make them as simple and direct as possible and, where feasible, consistent across test versions. The first time an item type was used in a version of the test, a model answer was provided so that candidates were aware of the type of response required. Examples of different item types were included in the candidate information booklet.

Summary

In spite of efforts to select reading texts and design items to reflect natural language interactions, the construct of reading which can be tested in a test such as access: is necessarily a much reduced one, because of the constraints of the conditions of test-taking and test-marking. While current views of reading pedagogy (Wallace 1992) favour an integration of attention to the language code (discourse, grammar and vocabulary choices), strategies for accessing meanings, and more recently to critical awareness at an ideological level, test items remain on the whole limited to the first of these areas. Moreover the readings of the texts which are valued tend to be those which reflect the socio-cultural perspective of the test developers (Hill and Parry 1992). While acknowledging these limitations, care was taken in the development of the reading tests to avoid any obvious bias and to provide the clearest possible directions and explanations of tasks. The candidate information booklet became an important source of information in this respect, illustrating for candidates the kinds of texts and item types they might encounter.

Considerations in test construction: the writing module

Text and topic selection

Many of the issues which emerged in the design of the reading test also applied to the development of the writing test modules. Choice of text types and topics, for example, presented similar problems of relevance across such a broad candidature. Here again, an attempt was made to avoid highly predictable or bland topics without at the same time favouring one kind of candidate over others in topic choice.

The specifications for the writing test identified three categories of genre or social purpose for writing tasks. These were:

- Category 1 establishing and maintaining interpersonal contacts, or self-expression;
- Category 2 giving and/or requesting information or explanation;
- Category 3 arguing or discussing an issue.

Three texts was the maximum number that could feasibly be required within the time frame allowed for the writing module. By requiring candidates to produce one text in each category, the overall assessment of writing performance could take into account writing for a variety of purposes and audiences, drawing differently on resources of grammar and discourse. The first category could include, for example, personal correspondence, personal histories, very short fictional stories, or descriptions of people, places or things often set in the context of a personal letter. The second might include instructional texts...
explaining regulations or procedures, requests for information or goods or services, memos, or contributions to a report or newsletter. These were most often set within an employment context. The third might require a letter to the editor, or to an interest group, or a short essay. Once again the categories provided guidance and scope for item writers, and also consistency across test versions.

**Task variability**

It was anticipated that Section 1 (requiring the production of a Category 1 text) would be considerably easier for the majority of candidates, and that the other tasks would be progressively more difficult. Analysis of the test results generally supported the predicted progression in difficulty, but the differences were not as significant as had been anticipated. Some tasks developed for Category 1 gave rise to responses which varied greatly in their degree of formality of expression. However this variation was accommodated in the marking by alerting raters to the fact that such variation was to be tolerated.

**Task design**

For each task a writing space was provided, with design features appropriate to the kind of text required; for example, a workplace memo was provided with a formatted memo page for one task. In some cases, such as those in which the candidate was asked to write a personal letter, an opening few lines of text were sometimes provided, along with an appropriate closing.

As with the reading module, the rubrics for each task specified the approximate time that should be taken on that task and the required word length. The challenge in writing the rubrics for the writing tasks was to provide a writing prompt with sufficient but not too much information. Sufficient information had to be given in terms of context (purpose, audience, topic, text type, etc) to establish the requirements of the task, and to avoid the possibility of candidates providing rehearsed texts. At the same time, too much information could provide the candidate with significant amounts of language to transfer to their writing and thus invalidate the task. This was of particular concern as the allowable word length for each task was relatively short. In some cases, illustrations assisted in providing further information on a topic.

As a first check, item writers were required to undertake themselves the tasks they had designed and to submit their written response along with the task for review by the test development team. Other writers were also asked to do the tasks. This procedure was intended as a pre-trialling check to establish whether the tasks generated the kind of texts expected and whether the task could feasibly be completed within the word length limits. In the development process some adjustments were made to allowable word lengths to make them more realistic for the kinds of texts to be included and for the time frames of the module. Each version of the test had consistent requirements in terms of word lengths for each section. Sections 1 and 2 required approximately 150 words each and Section 3 required a text of between 150 and 200 words.

**Developing rating criteria**

The most challenging aspect of producing the writing test specifications was the development of assessment criteria and the accompanying descriptors for each level. As a starting point, the test development team drew on writing assessment descriptors developed for adult ESL learners in Australia (Hood and Navara 1991) as well as on descriptors used in other international English language tests (Westaway 1988; Hamp-Lyons 1990). Throughout the process of development, draft versions of criteria and descriptors were submitted to close review. A number of steps were taken to investigate the extent to which the descriptors reflected expert judgments of different levels of writing performance. These included asking a group of experienced assessors to match a jumbled set of descriptors to the appropriate criterion and level. The descriptors were also scrutinised by the TDC and reviewed on the basis of feedback from marking of trial data. A validation study of the writing performance descriptors is described in Chapter 9.

Four criteria were used in the final version of the specifications. These were:

- task fulfilment and appropriacy;
- conventions of presentation;
- cohesion and organisation;
- grammatical control.

Raters, who were all highly experienced ESL teachers, were given extensive training, including a trial marking and analysis of their relative leniency or severity and their consistency. A benchmarking activity also preceded each test-marking session to enable raters to rate and discuss scripts at different ability...
levels and to ensure consistency in the interpretation of the rating criteria. The FACETS statistical program was used to analyse the writing test results and to identify misfitting or inconsistent raters. (FACETS is described in the section ‘Test analysis’ in ‘Considerations in test construction: the oral interaction module’ earlier in this chapter). All writing performances were randomly assigned to two raters for independent double rating.

Summary

In a similar way to the reading test, the writing test takes a ‘product’ rather than a ‘process’ orientation to the construct of writing. That is, what is assessed is the outcome of the task, rather than the approach or strategies employed by the candidate. This seems to be an inevitable response to the operational constraints of test-taking and marking. While pedagogically it may be more appropriate to focus on writing as a collaborative social process, under test conditions the candidate is required to produce, unaided, a product of defined purpose and word length within a given time in order to enable reliable comparisons of performance. The writing tasks in the access: test attempt to give candidates clear information on the context, purpose and nature of the product required across a variety of genres. Again the candidate test information booklet is an important resource in helping to ensure that candidates are aware of the requirements of the test.

Conclusion

This chapter has aimed to present a picture of the evolution of the access: test and to describe some of the major challenges which arose during the test development process. As with any large scale project requiring test delivery in distant and sometimes under-resourced testing sites, the final product inevitably represented a compromise between, on the one hand, a desire to incorporate the best of current knowledge of the nature of language ability and, on the other, a need to work within the constraints imposed by time, resources and the limitations of standardised testing formats. Such compromises, however, are rarely easy and highlight the ‘continuous need [for language testers] to improve their procedures and products’ (Alderson et al 1995: 259). In this regard, it is hoped that the program of access: research inaugurated along with the test and reported in the remainder of this volume constitutes a solid foundation for such improvement.

Notes

1. For explanation of the terms ‘functional’ and ‘vocational’ see the section ‘Describing proficiency levels’ on p. 35.
2. Now the Department of Immigration and Multicultural Affairs.
3. Now Language Australia.
4. The Rasch model is one of a family of techniques known as ‘latent trait theory’ or ‘item response theory’. It allows testers to estimate the chances of a candidate of a given ability answering correctly an item of a given difficulty. One of the advantages of using this type of analysis is that it allows candidate ability and item difficulty to be estimated independently and reported on a common scale, thus avoiding some of the problems associated with sample-dependent classical measurement techniques. (For a fuller exploration of the Rasch model see McNamara 1996.)

References


3
Investigating second language listening ability: listening skills and item difficulty

Geoff Brindley

Introduction

Because listening comprehension is an invisible cognitive operation, it is very difficult to describe and hence to assess. This difficulty is reflected in a number of recent overviews of second language listening which conclude that a great deal of work still remains to be done to identify and describe models of listening comprehension which could be used to guide teaching and testing (Brown and Yule 1983; Rost 1990; Buck 1990, 1991a, 1992; Dunkel 1991, Dunkel et al 1992, Brindley and Nunan 1992).

Despite the lack of adequate theories and models of listening, numerous attempts — in the form of skills taxonomies, rating scales or band descriptors — have been made to describe listening ability. Many of these descriptions distinguish between different levels of ability on the basis of a presumed skills hierarchy and/or through a description of the kinds of tasks considered appropriate for different levels of listening ability.

Scale descriptors and skills taxonomies often form the basis of test or assessment task construction and in some cases may end up being accepted as a de facto definition of the listening construct itself (Davies 1992; North 1993). However, without independent evidence, we do not know that the skills claimed by the test developers to be being tested are in fact those that are tested. Thus, it is important to explore the relationship between skills taxonomies and test items by examining whether the skills described are reflected in test performance.

In order to address this question, this chapter investigates the extent to which expert judges are able to match particular listening skills to items and to agree
on item difficulty. In this case the taxonomy will be that used as the basis of the specifications for the listening component of the access: test.

**Issues in the assessment of listening**

**Inadequacies of current theories and models of listening**

The lack of adequate theories of listening comprehension which might underpin assessment is highlighted by Brown and Yule (1983: 100–101):

*We make it clear now that we find existing approaches to the assessment of listening comprehension based on a very insecure theoretical notion of what ‘comprehension’ means. It is by no means clear that a great deal of what is currently tested in listening comprehension tests is necessary, or relevant to the process of understanding the communicative event the student has listened to … Basic research on listening comprehension in extended texts is urgently required before any current methods of assessment can be regarded as appropriate diagnostic instruments and, in particular, before any current testing methodology can be regarded as providing a satisfactory motivation for constructing courses to teach listening comprehension in one way rather than another.*

This is not to suggest, however, that research has provided no guidance whatsoever to teachers and test developers or that there is no consensus on some basic theoretical issues. Buck (1990), in a review of the considerable literature on first and second language comprehension, suggests that the view of listening as a two-stage process seems to be very widely reflected in the research literature and comments that this seems to be intuitively accepted by teachers. The two stages, as described by Carroll (1972), are:

1. apprehending linguistic information (lower level processing);
2. relating the information to a broader context (macro-comprehension).

**Processing in second language listening**

These two hypothesised stages in comprehension raise the question of how language input is actually processed by the listener. In this context, the distinction between ‘top-down’ and ‘bottom-up’ processing, which figures in many accounts of listening, has been the subject of a great deal of debate in the literature. In bottom-up processing, the smallest units of language are identified first, and these are ‘chained together’ to form the next highest unit, these are chained together to form the next highest unit and so on. Top-down processing, on the other hand, depends on the use of context and background knowledge to understand the meaning of the incoming message.

In the context of second language comprehension, Kelly (1991) hypothesises that in the early stages of language learning, learners rely heavily on bottom-up processing, only bringing semantic and other knowledge into play at later stages. Richards (1988), however, argues that the opposite is the case, since learners at lower proficiency levels lack the linguistic knowledge to engage in data-driven processing. Richards’ explanation, according to Buck (1990: 81) seems more intuitively reasonable. (It is also implicitly supported in the descriptions of lower levels of listening ability in many proficiency scales which accentuate the learner’s dependence on contextual support and world knowledge in order to interpret aural texts).

In recent years, simple bottom-up or top-down explanations have been rejected as inadequate for explaining how second language learners process input. They have been replaced by interactive-compensatory models which are based on the view that, in processing discourse, information from more than one level is utilised simultaneously. Stanovich (1980) suggests that poor language users may actually be more dependent on higher level processes than proficient users, a suggestion which is consistent with several empirical studies and with the views of Richards (1983) cited above. In the context of listening comprehension, however, the fact that a listener may activate higher level and lower level processes simultaneously suggests that it would be extremely difficult to unequivocally link particular skills and processing strategies to the interpretation of a given aural text.

**Factors affecting second language listening**

A wide range of variables affect the processing demands of listening tasks and, by extension, of listening test items. Termed ‘levelling variables’ by Dunkel, Chaudron and Henning (1993: 184), Anderson and Lynch (1988) describe these variables as being similar to volume controls on a radio. Any item may be affected by the ‘volume setting’ of one or more of these variables. While factors such as background knowledge, speech rate, and syntactic complexity are acknowledged in the literature on listening materials design as crucial factors which affect task difficulty, there have been few attempts to systematically incorporate...
these variables into test design, and little is known about their relative effects on levels of comprehension. This makes it very difficult to explain why certain items should be more difficult than others. A number of particularly important factors, however, have been isolated which might serve as a basis on which to differentiate levels of ability and/or task difficulty. These include the following.

**Lexical knowledge**

Lexical knowledge is clearly important in listening; lack of knowledge of key lexis can lead to miscommunication or even breakdown. The study conducted by Boyle (1984: 37) into student and teacher perceptions of key factors in comprehension found that students gave a good deal more emphasis to the importance of vocabulary than the teachers did. This confirms the view of Kelly (1991) who found that ignorance of vocabulary was the major factor contributing to lack of listening comprehension beyond the intermediate level of language learning.

**Syntactic knowledge**

In a review of the role of memory in listening comprehension, Call (1985: 769) concludes that ‘target language syntax seems to be an important factor in increasing the amount of linguistic material that can be retained in short-term memory’.

Second language acquisition (SLA) research has also shown that ability to recognise form-function relationships is an important element in the development of proficiency, which of course includes listening (Brindley 1987). In fact, SLA has been characterised as a process whereby ‘loose paratactic, “pragmatic” discourse structures develop over time into tight “grammaticalised” syntactic structures’ (Givon 1979: 208). This description applies just as much to the development of listening as it does to the development of speaking. ‘Receptive’ syntactic knowledge is clearly an important factor in listening ability.

**Background knowledge**

The central insight provided by researchers using mental models such as frame and schema theory is that meaning does not come neatly prepackaged in aural and written texts. Widdowson (1978) has suggested that texts are little more than elaborate ‘signposts’ to the speaker or writer’s original meanings, and that the reader or listener must use their linguistic and content knowledge to reconstruct the original meanings of the creator of the discourse. Clearly what is inside the listener’s head is of major importance in determining how an aural message will be interpreted.

Summing up the role of scripts, frames and schemata in listening comprehension, Buck (1990: 71) states that:

> it is clear that background knowledge, whether conceived as frames, scripts or schema, is obviously important in listening comprehension. If the listener shares the same knowledge as the speaker, much of what is being said can be understood in terms of top-down constraints. However, if the listener has no knowledge relevant to a particular event these top-down constraints will be lacking and comprehension will be more dependent on lower-level information. In cases where the listener’s background knowledge for a specific event is different from the speaker, not only may helpful top-down constraints be lacking, but there is the possibility of the listener being led down various garden paths in an attempt to understand events in terms of inappropriate knowledge structures.

It seems reasonable to assume, as Brown and Yule (1983: 84) suggest, that the more background knowledge which is taken for granted in a particular discourse, the more difficult it will be for a learner who does not share that knowledge.

**Speech rate**

Buck (1990: 91–93) reviews a range of studies on the effect of speech rate on second language comprehension, all of which suggest that the faster rates of delivery can significantly reduce comprehension. He concludes that:

> the rate of delivery is a very important variable in the comprehension of a second language; a fact which accords well with the personal experience of most teachers and second language learners.

**Noise**

The ability to understand messages in adverse conditions is one factor which appears to differentiate native speakers from non-natives. Kelly (1991: 138), for example, quotes a study by Whiteson (1972) which compared the auditory
Describing hierarchies of listening skills: issues and problems

Despite the uncertainty surrounding the nature of listening skills and their hierarchical relationship, numerous rating scales and taxonomies purporting to describe levels of listening ability have been produced (see Brindley and Nunan 1992 for a review). However, most of these appear to be based on experience and intuition rather than on research, and it is clear that further research studies are required to establish the extent to which they represent the processes and strategies that second language listeners actually deploy (Buck 1990; Brindley and Nunan 1992).

Brindley and Nunan (1992) have raised a number of concerns in relation to validation studies of rating scales which are carried out through parallel test construction. One is the essential circularity of attempting to validate proficiency descriptors through constructing tasks or tests based on those descriptors. This is illustrated by a study of tests based on the ACTFL scales conducted by Dandonoli and Henning (1991) in which the researchers were unable to conclude whether the failure of some of the test results to conform to the predicted task/text hierarchies was because the tests were poorly constructed or because the descriptors themselves were inadequate (elsewhere, however, Dunkel et al. 1993 seem to be of the view that if the tests demonstrate high internal reliability, then the fault may lie with the scale descriptors). These difficulties are further compounded by the fact that it is hard to see how the kind of impressionistic terminology which is typically used in rating scales such as the ACTFL scale could be translated directly into test specifications or items — subjective judgement would still have to play a major role. On what basis, for example, would test designers decide what is a ‘linguistically noncomplex text’ (ACTFL Intermediate-Mid Reading) or write items testing ‘emerging awareness of culturally implied meanings beyond the surface meanings of the text’ (ACTFL Advanced-Plus Listening)?

Unless there is independent evidence to demonstrate that candidates at particular levels can indeed handle the tasks and texts and demonstrate the skills specified as accompanying that level, it cannot be assumed that the verbal descriptions provided in rating scales describe the skills used in test performance. As Alderson (1991b) points out, if descriptors are to be meaningful characterisations of ability, then they should be able to be related to actual performance.
Findings such as these would seem to suggest that the descriptions of skills hierarchies which figure in listening rating scales and skills taxonomies are open to question, firstly because of the lack of evidence to support the existence of separately identifiable subskills, and secondly because of the failure of studies such as those cited to show implicational relationships between skills.

It should be noted, however, that the methodology used by Alderson (1990a) has been criticised by Weir, Hughes and Porter (1990) who argue, *inter alia*, that the key notions of ‘higher order’ and ‘lower order’ skills are inadequately defined in the study. Weir et al. also contend that since Alderson did not train judges to make reliable judgements, their lack of agreement was hardly surprising. Lumley (1993), in a partial replication of Alderson’s study, reports that following intensive briefing and discussion, he was able to achieve substantial levels of agreement among judges who were asked to rate the difficulty of subskills in an EAP reading test. Alderson (1991a), however, in a reply to Weir et al., argues that training judges is inappropriate if one wishes to establish the extent to which they share a common understanding of skills and item difficulty. If judges disagree, he points out (Alderson 1991a: 600), then:

> this surely tells the researcher something important about their judgements or their understanding of the concepts on which the judgements are based.

The study described in this chapter is intended as a contribution to the line of research undertaken by Alderson, Buck and others into the way that listening skills are described, operationalised and interpreted by test constructors and users. It is hoped that the results will be of some assistance in clarifying the complex relationship between skills specifications and actual test performance and will suggest further avenues for investigation of the interaction between listening processes and test tasks.

**Research questions**

The study set out to address the following research questions:

- To what extent do experienced language teachers agree on the listening skills being tapped by particular items?
ensuing discussion, there were several suggestions concerning the ordering of the skills and a number of changes were made, such as the addition of 'explicitly stated in text' in skills 2.3 and 3.1. The revised taxonomy appears in Table 1.

A piloted version of the access: listening skills test was then administered to the group. The test consisted of six listening passages in five sections, comprising a total of 43 items. The item formats are listed in Table 2. Group members were then asked to:

- identify, using the agreed taxonomy, the skill or skills that they thought each item was testing and to specify, if they felt they were able, the essential skill involved;
- rate the items from easiest to most difficult on a scale of 1 to 3.

### Table 1: Revised listening skills taxonomy (adapted from Weir 1993; Rost 1990)

<table>
<thead>
<tr>
<th>1</th>
<th>Orienting oneself to a spoken text</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Identifying the purpose/genre of a spoken text</td>
</tr>
<tr>
<td>1.2</td>
<td>Identifying the topic</td>
</tr>
<tr>
<td>1.3</td>
<td>Identifying the broad roles and relationships of the participants (eg superior/subordinate)</td>
</tr>
<tr>
<td>2</td>
<td>Identifying the main idea/s in a spoken text</td>
</tr>
<tr>
<td>2.1</td>
<td>Distinguishing main ideas from supporting detail</td>
</tr>
<tr>
<td>2.2</td>
<td>Distinguishing fact from example</td>
</tr>
<tr>
<td>2.3</td>
<td>Distinguishing fact from opinion when explicitly stated in text.</td>
</tr>
<tr>
<td>3</td>
<td>Extracting specific information from a spoken text</td>
</tr>
<tr>
<td>3.1</td>
<td>Extracting key details explicitly stated in text.</td>
</tr>
<tr>
<td>3.2</td>
<td>Identifying key vocabulary items</td>
</tr>
<tr>
<td>4</td>
<td>Understanding discourse structure and organisation</td>
</tr>
<tr>
<td>4.1</td>
<td>Following discourse structure</td>
</tr>
<tr>
<td>4.2</td>
<td>Identifying key discourse/cohesive markers</td>
</tr>
<tr>
<td>4.3</td>
<td>Tracing the development of an argument</td>
</tr>
<tr>
<td>5</td>
<td>Understanding meaning not explicitly stated</td>
</tr>
<tr>
<td>5.1</td>
<td>Relating utterances to the social/situational context</td>
</tr>
<tr>
<td>5.2</td>
<td>Identifying speaker’s attitudes/emotional state</td>
</tr>
<tr>
<td>5.3</td>
<td>Recognising the communicative function of stress/intonation patterns</td>
</tr>
<tr>
<td>5.4</td>
<td>Recognising the speaker’s illocutionary intent</td>
</tr>
<tr>
<td>5.5</td>
<td>Deducing meaning of unfamiliar words</td>
</tr>
<tr>
<td>5.6</td>
<td>Evaluating the adequacy of the information provided</td>
</tr>
<tr>
<td>5.7</td>
<td>Using information from the discourse to make a reasonable prediction</td>
</tr>
</tbody>
</table>

In order to obtain an overview of the extent to which judges were able to agree on the assignment of skills to items, all items were tabulated along with the skills assigned to it by each judge. A summary of the range of skills assigned to the ten easiest and most difficult items is given in Table 3.

At the same time, item analysis of the results of an administration of the same test to 256 adult non-native speakers of English was carried out using the Rasch-based QUEST program (Adams and Khoo 1992). This program gives item difficulty estimates in the form of logits, as well as other classical indices such as point-biserial correlations and facility indices.
who thought that it was the main skill being tested. It can be seen that each item was attributed to a range of skills (cf Alderson 1990a); in no case was there complete agreement that an item was tapping only one skill. However, there was some agreement on particular item types, for example items 1, 3 and 4 which required testees to transcode specific details from a telephone conversation. Similarly, Items 27 and 29 demanded the transcription of numbers from a conversation. Those item types on which there was less consensus tended to be the multiple choice questions (see, for example, items 40, 41 and 42, where judges reported that there were several ways of arriving at the answer, and the gapped summary (see item 33).

### Item difficulty

Judges' ratings of item difficulty were correlated with each other and with the logit estimates of difficulty obtained in the item analysis. The correlation matrix is provided in Table 4 on p. 78.

Although there were moderate to high correlations between some pairs of raters — ranging from .61 to .92 — only one of the judges' ratings correlated significantly — at .51 — with the item difficulty estimates. This judge was, interestingly, an item writer and had not at that stage seen the results of the item analysis. The ratings of the seven raters were also pooled and the correlation of the average rating with the item difficulty estimates calculated. This correlation was .31, which was significant at .05, but nevertheless indicative of a weak relationship between judges' average ratings of item difficulty and the results of the item analysis.

### Discussion

Before considering the significance of the outcomes of the study, some methodological issues need to be clarified in order to place the discussion of results in context and at the same time to point out the limitations of the study.

Firstly, it should be emphasised that, although the listening skills taxonomy was discussed and exemplified at the outset, no attempt was made to train raters to agree, as might be done if they were being required to rate oral language. Although it could be claimed that this lack of training may have contributed to the lack of agreement between judges, it should be emphasised that one of the two main aims of the study was to investigate individual perceptions of how the listening skills described in the taxonomy were reflected in test items and to see

### Table 3: access: listening module: Item difficulties and skills ratings (n=256)

<table>
<thead>
<tr>
<th>Item</th>
<th>Logit diff</th>
<th>Mean rating</th>
<th>Facility value</th>
<th>Pt bis</th>
<th>Major skill(s) assigned by judges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harder</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2.12</td>
<td>1.64</td>
<td>39.9</td>
<td>.30</td>
<td>3.1(1), 3.2(5)</td>
</tr>
<tr>
<td>30</td>
<td>1.98</td>
<td>1.71</td>
<td>24.6</td>
<td>.54</td>
<td>2.1(2), 3.1(2), 3.2(2), 4.1(1)</td>
</tr>
<tr>
<td>33</td>
<td>1.97</td>
<td>2.96</td>
<td>31.1</td>
<td>.41</td>
<td>2.1(3), 3.1(3), 4.1(1), 4.2(1), 4.3(1)</td>
</tr>
<tr>
<td>4</td>
<td>1.89</td>
<td>1.6</td>
<td>29.8</td>
<td>.34</td>
<td>3.1(2), 3.2(6)</td>
</tr>
<tr>
<td>8</td>
<td>1.70</td>
<td>1.9</td>
<td>32.0</td>
<td>.48</td>
<td>2.1(2), 3.1(3), 3.2(3)</td>
</tr>
<tr>
<td>40</td>
<td>1.66</td>
<td>3.0</td>
<td>30.6</td>
<td>.38</td>
<td>2.1(1), 4.3(1), 5.1(3), 5.5(1)</td>
</tr>
<tr>
<td>41</td>
<td>1.63</td>
<td>2.4</td>
<td>30.8</td>
<td>.50</td>
<td>2.1(2), 3.2(2), 4.1(2), 5.1(1)</td>
</tr>
<tr>
<td>42</td>
<td>1.40</td>
<td>2.6</td>
<td>34.6</td>
<td>.40</td>
<td>2.1(2), 4.1(1), 4.3(1), 5.1(2)</td>
</tr>
<tr>
<td>1</td>
<td>1.33</td>
<td>1.5</td>
<td>39.9</td>
<td>.51</td>
<td>3.1(1), 3.2(7)</td>
</tr>
<tr>
<td>28</td>
<td>1.29</td>
<td>1.7</td>
<td>41.1</td>
<td>.40</td>
<td>2.1(2), 3.1(3), 3.2(1), 4.1(1)</td>
</tr>
<tr>
<td>Easier</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>-3.24</td>
<td>1.9</td>
<td>96.8</td>
<td>.25</td>
<td>2.1 (1), 3.1 (2), 5.0 (2)</td>
</tr>
<tr>
<td>18</td>
<td>-2.71</td>
<td>1.9</td>
<td>94.2</td>
<td>.22</td>
<td>2.1 (1), 3.1 (1), 3.2 (3), 4.3 (1)</td>
</tr>
<tr>
<td>27</td>
<td>-2.63</td>
<td>1.1</td>
<td>93.6</td>
<td>.31</td>
<td>3.1 (1), 3.2 (6)</td>
</tr>
<tr>
<td>32</td>
<td>-2.41</td>
<td>2.1</td>
<td>91.1</td>
<td>.35</td>
<td>2.1 (1), 3.1 (3), 3.2 (3)</td>
</tr>
<tr>
<td>16</td>
<td>-1.31</td>
<td>1.9</td>
<td>82.4</td>
<td>.44</td>
<td>2.1 (1), 3.1 (1), 3.2 (2), 4.3 (3)</td>
</tr>
<tr>
<td>43</td>
<td>-1.35</td>
<td>2.3</td>
<td>81.7</td>
<td>.46</td>
<td>2.1 (3), 2.3 (1), 4.1 (1)</td>
</tr>
<tr>
<td>38</td>
<td>-1.25</td>
<td>2.3</td>
<td>80.5</td>
<td>.46</td>
<td>2.1 (2), 3.2 (2), 4.3 (1), 5.1 (1)</td>
</tr>
<tr>
<td>10</td>
<td>-1.02</td>
<td>1.2</td>
<td>76.7</td>
<td>.45</td>
<td>3.1 (3), 3.2 (4)</td>
</tr>
<tr>
<td>29</td>
<td>-1.02</td>
<td>1.1</td>
<td>77.4</td>
<td>.42</td>
<td>3.1 (3), 3.2 (4)</td>
</tr>
<tr>
<td>31</td>
<td>-0.49</td>
<td>1.5</td>
<td>74.6</td>
<td>.38</td>
<td>2.1 (4), 3.1 (2), 3.2 (1)</td>
</tr>
</tbody>
</table>

Judges' estimates of item difficulty were then correlated with the Quest-generated estimates of item difficulties from the test administration. The descriptive statistics for the pilot test were as follows:

Mean: 25  
SD: 11.8  
Rasch reliability of case estimate (Rasch equivalent of KR20): .89

### Results

#### Item/skill matching

The last column in Table 3 shows the skills attributed to the ten easiest and most difficult items. The numbers after each skill indicate the number of judges who thought that it was the main skill being tested. It can be seen that each item was attributed to a range of skills (cf Alderson 1990a); in no case was there complete agreement that an item was tapping only one skill. However, there was some agreement on particular item types, for example items 1, 3 and 4 which required testees to transcode specific details from a telephone conversation. Similarly, Items 27 and 29 demanded the transcription of numbers from a conversation. Those item types on which there was less consensus tended to be the multiple choice questions (see, for example, items 40, 41 and 42, where judges reported that there were several ways of arriving at the answer, and the gapped summary (see item 33).

### Item difficulty

Judges' ratings of item difficulty were correlated with each other and with the logit estimates of difficulty obtained in the item analysis. The correlation matrix is provided in Table 4 on p. 78.

Although there were moderate to high correlations between some pairs of raters — ranging from .61 to .92 — only one of the judges' ratings correlated significantly — at .51 — with the item difficulty estimates. This judge was, interestingly, an item writer and had not at that stage seen the results of the item analysis. The ratings of the seven raters were also pooled and the correlation of the average rating with the item difficulty estimates calculated. This correlation was .31, which was significant at .05, but nevertheless indicative of a weak relationship between judges' average ratings of item difficulty and the results of the item analysis.
to what extent these perceptions coincided with those of other judges, rather than to obtain group consensus.

Related to the issue of rater agreement is the role of skills taxonomies in research of the kind described in this chapter. In cases where an investigator wants solely to probe people’s understanding of the specific skills required to answer items, an existing taxonomy might not be the most appropriate starting point as it could restrict judges’ attention to the skills on the list. In such cases, it would be useful to explore the use of more open-ended techniques whereby judges are invited to introspect on the skills being tapped by test tasks, without providing them with an existing taxonomy.

Secondly, a note of caution should be sounded with respect to the intercorrelations between judges’ ratings of item difficulty. This is because three of the judges allocated difficulty levels to entire sections rather than to individual items as requested. This was probably due to time constraints, but undoubtedly affected the correlation with the item difficulty ratings. Closer attention to individual items would have led to more variation in difficulty ratings within sections.

Thirdly, it should be emphasised that the judges, although requested to identify the key skill being tested by each item, were extremely reluctant to assign a single skill to an item. In the discussion preceding the rating exercise, they pointed out that an item may require a number of skills to be used simultaneously. This means that the ‘most important skill’ rating needs to be treated with caution.

These reservations notwithstanding, a number of remarks can be made about the results of this study. Firstly, the failure of judges to agree on the assignment of skills can be attributed to a variety of factors. As already noted, judges felt that items often required several skills and this is consistent with current views of listening as a parallel and interactive process, and with ‘the common-sense view that answering a question is likely to involve a variety of interrelated skills, rather than one skill only, or even mainly’ (Alderson 1990a: 436).

However, the possibility of agreement may also have been lessened by the way in which the skills themselves were described; that is, independently of the test-taking processes. The skills taxonomy used did not make specific reference to item format or response mode which forced the judges to focus more on the listening skills in the abstract than on what was happening as part of test-taking specifically. For this reason, it may have been preferable to have developed a

| Table 4: Correlation matrix of judges’ skills ratings and QUEST item difficulty estimates |
| ID | R1 | R2 | R3 | R4 | R5 | R6 | R7 | TOTR |
| ID | 1.00 | 0.214 | 0.264 | 0.428** | 0.925** | 0.480* | 0.332* | 0.310* |
| R1 | 1.00 | 0.925** | 0.421* | 0.438* | 0.224 |
| R2 | 1.00 | 0.925** | 0.421* | 0.438* |
| R3 | 0.224 | 0.264 | 0.428** | 0.925** |
| R4 | 0.510** | 0.221 | 0.144 |
| R5 | 0.626** | 0.144 |
| R6 | 0.612** | 0.657* |
| R7 | 0.626** | 0.657* |
| TOTR | 0.310* | 0.787** | 0.667* | 0.849* |

ID = Item difficulty  
R1 = Rater 1 etc  
TOTR = Total ratings  
* = Correlation significant at p<.05  
** = Correlation significant at p<.01
In considering the question of item difficulty, it is worth bearing in mind, as Buck (1990) demonstrates, that each item is itself the product of multiple influences. Under actual test conditions it is thus likely that the 'difficulty' of the most difficult items is due to a combination of the characteristics of the input (eg speed, accent), the internal organisation of the test (eg the position of the item in the test) and/or the response mode (eg whether note-taking is required and if this requires synthesis or merely transcoding) rather than solely to the intrinsic difficulty of the listening skill involved. In this regard it is interesting to note that when Items 1 and 3 were re-recorded at a slower speed for inclusion in a new version of the test, they become noticeably easier. This highlights the need to investigate the effect of systematic manipulation of key variables such as those listed above in order to clarify the relative effects on item difficulty.

**Conclusion**

Like the studies conducted by Alderson (1990a) and Buck (1990, 1992) this study has failed to establish consistent agreement among judges as to the skills being tested by particular items. Moreover, although some agreement was found in perceptions of item difficulty, these perceptions were largely at odds with the way in which the items performed when the test was administered to a population of non-native speakers.

Some of the methodological factors outlined above may have contributed to these results. In particular, it could be argued that the failure to distinguish between listening skills and listening test-taking skills may have made it difficult for judges to assign skills to items. However, as the judges themselves pointed out, listening requires the simultaneous exercise of several skills. The fact that so many skills were attributed to the items may be simply a reflection of this view of listening. It may also be evidence of the fact that people approach test items in different ways (Alderson 1990b: 477). Therefore the search to assign single skills to items may be fruitless.

The discrepancy between judges’ perceptions of item difficulty and actual difficulty tempts one to conclude, as did Buck (1991b: 19) that:

```
most experts really do not know what makes one item more difficult than another, and thus have little or nothing to offer the researcher interested in isolating those factors which influence item difficulty.
```
Perhaps the best way to address the question of item difficulty is through a thorough examination of second language listening processes in operation, supplemented by intensive analysis of items grouped at similar difficulty levels, rather than through appeal to expert judgement.

Finally, from a practical test design point of view, it is unlikely that specifying individual listening skills *a priori* is of great use to writers of listening tests. Although little information seems to be available on how items are actually developed, experience would seem to indicate that test developers do not start with a list of skills and then proceed to create items aimed at tapping those specific skill. Rather, the starting point is a text which by definition will tap some of the skills listed. What skills taxonomies of the kind given in test specifications can do is to assist test constructors *a posteriori* to see how comprehensively the range of skills has been sampled and to make modifications and additions to items as necessary. Further investigation of listening processes in conjunction with test content analysis of the type suggested above will help to ensure that these taxonomies better reflect the realities of language use and thus make a contribution to more valid tests of listening comprehension.

**Notes**

1. Parts of this chapter draw on Brindley and Nunan (1992). The contribution of David Nunan to this paper is acknowledged.

**References**


4

Negotiated discourse and interlanguage accent effects on a second language listening test

Steven Ross and John Langille

Introduction

The testing of second language listening comprehension is commonly conducted under the assumption that the non-native listener is an auditor of the target language as it is spoken in the daily interaction of native speakers. Very few second language listening tests are designed with the assumption that the second language listener is a participant in the speech event. The vast majority of such tests are based on narratives, dialogues and naturalistic interactions among fluent users of the language, who rarely have cause to accommodate or modify the speed or articulation of their speech in order to facilitate uptake by their interlocutor. The task of candidate listeners in such tests is to monitor and comprehend the propositional content of these excerpts from real life, as if they were native speakers of the target language.

Yet many of the tasks set for second language comprehension do not in fact authentically transpire without extensive two-party negotiation, backchanneling, redundancy, and clarification. Perhaps with the exception of announcements of factual detail, such as those at railway stations or airports, most contexts for speaking and listening are interactive and accommodative to non-fluent listeners. The extent of interactivity is manifested in the control the non-native speakers can manage over the content, flow and redundancy of the interlocutor’s speech, commonly by signalling difficulty in the uptake of messages from the conversational partner.

Analyses of interactions between native speakers of a language and adult learners of that language consistently reveal a disparity between how language is used
natively and how it is accommodated to the learner. Key features of accommoda-
tion are linguistic simplification of the code (Tickoo 1993) and inter-
actional modification of the discourse in which it is embedded (Long 1983a, 1983b). Verbal interaction between native and non-native speakers of a language almost always contains considerable redundancy and paraphrase so as to increase the likelihood of comprehension. Furthermore, research suggests that the amount of redundancy in native speaker–non-native speaker interaction is a function of the anticipated facility of comprehension (Ross and Berwick 1992). Second language listening tests, in contrast, very rarely incorporate interactionally modified discourse into their design, although such modification may have considerable impact on task difficulty and content validity of the measurement instrument.

This gap between how second language tests are typically constructed and what second language listeners actually do in order to process contextualised two-
way interaction leads to a potential difficulty for the content and construct
validity of many, if not most, second language listening tests. However, the in-
tegration of authentic interaction into second language tasks opens a Pandora’s Box in terms of test method effects. A multiplicity of factors may affect listening
test performance. In addition to the latent schematic factors underpinning listening
tasks, such as background knowledge and presuppositions, there are the possible confounds of different accents, speech styles, and modes of interaction to complicate the picture. However, little is known about the specific effects of these factors on comprehension.

The present research focuses on two important influences on test difficulty —
the effect of negotiation on dialogues that contain novel lexical items, and the impact of non-native accents of participants in the interaction.

The study

Materials

Two parallel forms of a listening comprehension test were constructed for this study. Both forms contained two dialogues between a native speaker of Aus-
tralian English and an English as a second language speaker (Thai) interlocutor. The dialogues were scripted so they contained a comparable number of propo-
sitions and novel lexical items, then two versions of each dialogue were created. In the first version of each dialogue (Baseline) novel lexical items (see Table 1) occurred in the discourse without any redundancy or negotiation as to mean-
ings. The second version (Negotiated) contained an equal number of propo-
sitions and novel lexical items, but had redundancy embedded in the discourse through negotiation of the novel lexical items at junctures occurring immedi-
ately after their first mentionings in the discourse. These redundancies potentially function to make much of the propositional content of dialogue opaque to non-native listeners. An example of how the baseline and negotiated dialogue might differ can be seen in the following comparison.

Baseline

X: What do you do?
Y: I’m an ornithologist.
X: Sounds interesting.
Y: It’s for the birds.

Negotiated

X: What do you do?
Y: I’m an ornithologist.
X: What is that?
Y: I study birds.
X: Sounds interesting.
Y: It’s for the birds.

The two parallel versions of each dialogue were then recorded and became the basis of Test forms A and B with Test form A containing the baseline version of Dialogue 1 and the negotiated version of Dialogue 2 and Test form B containing Baseline dialogue 2 and Negotiated Dialogue 1 (see Appendices 1 and 2).

Table 1: Novel lexical items in study dialogues

<table>
<thead>
<tr>
<th>Novel lexical items in study dialogues</th>
</tr>
</thead>
<tbody>
<tr>
<td>civil engineer</td>
</tr>
<tr>
<td>aviation engineer</td>
</tr>
<tr>
<td>geophysics</td>
</tr>
<tr>
<td>aerodynamics</td>
</tr>
<tr>
<td>seismology</td>
</tr>
<tr>
<td>apprentice</td>
</tr>
<tr>
<td>biochemistry</td>
</tr>
<tr>
<td>neonatologist</td>
</tr>
<tr>
<td>pathologist</td>
</tr>
</tbody>
</table>

Inclusion of accents

Since the key propositions and novel lexical items were uttered in the dialogues by one or the other of the interlocutors, the potential effect of the accent of the speaker can be assessed by examining relative differences in the difficulty of
propositions and lexical items referred to by either the native or non-native speaker in each dialogue. The two Thai speakers participating in the study contribute a non-native accent that is characterised by nasalisation of unstressed vocalic segments and tonic stress placement quite different from their Australian interlocutors. The inclusion of non-native accents in the study facilitates an analysis of the impact of interphonology on the comprehensibility of baseline and negotiated dialogue versions contributed by the Thai speakers. Specific propositions and lexical items uttered by the Thai native speakers constitute a subset of the total items on the listening task, and thus allow for an examination of the impact of non-native accent on comprehensibility. This comparison is achieved by noting listener comprehension (manifested in item difficulties) of propositions uttered by the Thais compared with those uttered by Australian English native speakers.

Method

Tapes and test forms were distributed in a quasi-experimental design, which was devised to lend itself to an internal replication of the main effects analysis. Since, for the most part, intact groups were used within institutions participating in the project, and no scheme for thorough randomisation could be utilised, a two-part dictation was given immediately prior to and after the two main listening tasks. The dictations were designed to serve as the basis for examining the comparability of the groupings of listeners, and when summed to serve as a covariate in the event the groups are not equivalent in listening proficiency at the outset. They also provided the basis for matching subjects for more detailed analyses of accent comprehension differences.

Data was collected in Australia, Finland and Japan. In Japan, at each institution where the quasi-experiment was conducted, two roughly parallel groups of learners were recruited. Instructors were asked to identify two approximately equal groups of students in the same course and then to arbitrarily assign Form A to one group and Form B to the other. In Finland, parallel groups of learners were recruited on the basis of the number of years of English language study using beginning-intermediate-advanced level criteria. The beginning and intermediate groups were taken from sixth and ninth forms of high schools respectively and the advanced groups consisted of undergraduates or graduates who had majored in the English language at various tertiary institutions throughout the country. As in Japan, the instructor was unaware of the purpose of the research and assigned the groups either Form A or B on a random basis. Schematically, the administration of the materials were as in Table 2.

<table>
<thead>
<tr>
<th>Form A</th>
<th>Dictation 1</th>
<th>Dialogue 1 Baseline</th>
<th>Dialogue 2 Negotiated</th>
<th>Dictation 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Form B</td>
<td>Dictation 1</td>
<td>Dialogue 2 Baseline</td>
<td>Dialogue 1 Negotiated</td>
<td>Dictation 2</td>
</tr>
</tbody>
</table>

The recordings were designed so that the same dialogues appeared on both forms in either the baseline or negotiated version. The Form A Baseline dialogue, for example, appears again in Form B in the Negotiated format. Thus, different groups of listeners were exposed to either the baseline or negotiated forms, in order to contrast the impact of baseline versus negotiated discourse forms of the listening task.

Since there were two parallel sets of dialogues, an internal replication of main effects was possible in this design. Both the main study and the replication compared the effect of the baseline version of the dialogue with its negotiated counterpart. Henceforth, the main study, which contrasts the baseline and negotiated versions, will be referred to as the ‘source’ analysis, and the second study, conducted within the same institution with parallel groups of listeners, will be referred to as the ‘replication’ study. The purpose of the replication was to strengthen the basis for generalisation about whatever main effects were observed in the source study by providing empirical evidence for their reproducibility with parallel groups of listeners.

Procedures

Listeners were instructed to take notes in the box provided on the answer sheet and then to write a summary of the dialogues in their native language beneath the box (see Appendix 1 for an example in English). These first language (L1) summaries were subsequently scored by awarding one point for each accurate citation of a proposition appearing in the dialogue (thirty propositions in each form of the dialogues were identified). This scoring scheme allowed for analysis of the comprehension of the overall propositional content of the dialogues as well as comprehension of the novel lexical items mentioned and defined therein.
Participants

In order to determine the difficulty of each proposition, a group of fifty Australian undergraduates and university staff completed one or the other form of the listening and summary tasks. The responses by Australian native speakers of English serve as the starting point for determining the salience and citability of each proposition, as well as a basis for determining the difficulty of the novel lexical content cited in the dialogues.

In addition to the 50 Australian native speakers of English, 314 Finnish and 274 Japanese subjects participated in the study. All of the non-native subjects were either enrolled in intact English as a foreign language classes at the high school/college levels or were English major undergraduates/graduates of tertiary institutions in their respective countries. An additional focus of the study might have been the difference between Australians, Finns and Japanese on the comprehension of the dialogues. However, since the sampling strategies differed in the two countries, such a comparison would not be enlightening. In Finland, a random sampling was conducted at the beginning, intermediate and advanced levels (determined by the number of years of English language study: 6, 9, 9+, respectively). In contrast, the sampling conducted in Japan was mainly from narrowly homogeneous groups of first-year university students. The differences in proficiency between the Finns and Japanese precludes useful comparison, although analysis of interactions between first language and comprehension remains viable. The main object of interest here is whether main effects hold across the three first-language groups featured in the study.

Hypotheses about negotiation

The primary hypothesis concerned the effect of negotiated discourse in the interactions between the Australians and their Thai interlocutors. If the redundancy provided by definitional statements provided as rejoinders to queries about statements made by the interlocutor did, in fact, serve to make propositional content more salient to listeners, the effect of such redundancy should be manifested in higher mean accuracy in the summaries written by the listeners. Since the design of the study employed an internal replication — essentially two parallel experiments, it was anticipated that the negotiated versions of the dialogues would be more salient than the unembellished baseline versions on both forms distributed.

The first hypothesis was that on Form A of the test, the summaries for Dialogue 1 (Baseline) would show a lower mean proposition citation for all subjects than would Form B (Negotiated). In the replication study, the same prediction was made, but the opposite pattern of mean differences was expected; On Form B, the mean would be lower for Dialogue 2 (Baseline) summaries than for Dialogue 1 (Negotiated). These predictions were expected to hold across the three groups sampled.

Hypotheses about accent

The issue of non-native accent on second language listening tests is one that has not been extensively explored. The interrelatedness of negotiation and accent complexity is important for the identification and operationalisation of the construct of ‘international’ English, and crucially, for the testing of it (Bhatia 1995). An issue for the assessment of international English is the extent to which varying accents will affect the difficulty of test items. The study considered the difference in comprehensibility of two accents, Australian and Thai, and thus constitutes a fixed effect design (Jackson and Brashers 1994). Of larger importance was the question of the impact of the negotiation factor on the relative comprehensibility of each accent. It was hypothesised that phonologically complex accents might be modifiable with the use of discourse which includes some degree of negotiation. In any event, the negotiation factor also potentially enhances the authenticity of interlanguage discourse, since it affects the interaction between native and non-native speakers in a manner more akin to the type of interactional modification observed when interlocutors perceive difficulties in communication.

In order to assess the impact of the Thai accent on the listening tasks in this study, item analyses were performed on propositions cited in both baseline and negotiated forms of the task. If it were the case that the non-native accent systematically affected the comprehensibility of the utterances, then the difficulty of comprehending those utterances and reproducing the underlying propositions in the summaries would differ. Since the effect of negotiation would be controlled through counterbalancing, the expected differences would be attributable to accent differences between the Australian and Thai speakers.

For the item difficulty analysis, the one-parameter logistic (Rasch) model (Andrich 1988) was used to scale the items for matched groups of Japanese and Finnish non-native listeners. The basis for the matching was the total dictation score. The matching resulted in a stratified sampling of the Finnish and Japanese listeners with a non-significant mean difference between the forms
Factorial analysis of covariance was used to determine whether the main effects owing to the negotiation of discourse would be symmetrically observable across the three first-language groups. This analysis allowed for an examination of interactions between the form (negotiation or baseline) and the native language of the listeners. The two covariates function to diminish residual error and compensate for small initial differences between the groups detected in spite of random assignment of group to treatment.

**Results**

The negotiation hypothesis was that summaries of Dialogue 1 (Baseline) in Form A would show a lower mean proposition citation for all subjects than would Dialogue 1 (Negotiated) on Form B. In the replication, the opposite order was predicted because the order of presentation was reversed. The analysis of covariance reveals that after adjustment for small differences on the dictation covariate and replication set responses, the predicted order of difficulty is observable in the mean accuracy of citations scored in the summaries. Table 4 shows the observed main effects for the ‘source’ negotiation study.

<table>
<thead>
<tr>
<th>Table 4: Means and standard errors for ‘source’ study</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>-----------------</td>
</tr>
<tr>
<td>Form A: Dialogue 1 (Baseline)</td>
</tr>
<tr>
<td>Form B: Dialogue 1 (Negotiated)</td>
</tr>
</tbody>
</table>

The impact of the main effect is more complex than mean comparisons would indicate. The effect of negotiation is asymmetric across the three language groups analysed. Since the proficiency of the listeners varied from native competence through advanced (Finns) to low intermediate and false beginners (Japanese), the influence of the negotiation depends largely on the native language and proficiency level. As can be observed in Table 5, there is the hypothesised main effect for the form of the dialogue heard (baseline or negotiated) as well as an effect attributable to the native language of the listeners.

The main effect is mediated by native language, but nevertheless applies in a parallel manner. The impact of negotiated discourse even serves to aid the native

---

**Negotiation study**

**Analyses**

As we have noted, a quasi-experimental design was used instead of a true experimental design which requires complete randomisation of listeners to forms, since only intact groups were accessible. In the quasi-experimental design, the two-part dictation was used as a covariate. The major requirements for using the dictation as a covariate are that it is highly correlated with the dependent variable of interest (listening comprehension) and that it be reliable (Huitema 1980). The two-part dictation used in this study was suitable on both accounts. In addition to the dictation, the score of the listeners on the parallel form served as a second covariate. Thus, when the negotiated version contrast is tested, the score on the baseline for the individual is correlated with the citation summary frequency; when the main effects test for the baseline is conducted, the score on each listener’s alternate task serves as the second covariate. This strategy serves to increase the power of the analysis by reducing the error term (Cohen 1988). Table 3 summarises the split-half and KR-20 reliabilities for the dictation and both forms of the listening task.

**Table 3: Reliability estimates**

<table>
<thead>
<tr>
<th>Dictation</th>
<th>Form A</th>
<th>Form B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>.93</td>
<td>.93</td>
</tr>
<tr>
<td>Negotiated</td>
<td>.82</td>
<td>.82</td>
</tr>
<tr>
<td>Baseline</td>
<td>.80</td>
<td>.80</td>
</tr>
<tr>
<td>Negotiated</td>
<td>.82</td>
<td>.82</td>
</tr>
</tbody>
</table>

The impact of the main effect is more complex than mean comparisons would indicate. The effect of negotiation is asymmetric across the three language groups analysed. Since the proficiency of the listeners varied from native competence through advanced (Finns) to low intermediate and false beginners (Japanese), the influence of the negotiation depends largely on the native language and proficiency level. As can be observed in Table 5, there is the hypothesised main effect for the form of the dialogue heard (baseline or negotiated) as well as an effect attributable to the native language of the listeners.

The main effect is mediated by native language, but nevertheless applies in a parallel manner. The impact of negotiated discourse even serves to aid the native

---

(\( F = .48, P = .49, df = 586 \)). Each item was calibrated according to the difficulty listeners showed in producing citations in their summaries. Of the thirty-one propositions on Forms A and B, thirty were retained in this analysis, one having no variance. Two calibrations were performed on each item, depending on the test form (A or B) listeners heard. Items were hypothesised to vary according to the accent of the speaker. It was expected that the accent of the Thai speakers would be more difficult than that of the Australian speakers, regardless of the extent of negotiation manifested in the discourse. Here, a factorial repeated measures analysis of variance was used to examine the variance attributable to speaker and form version.
speakers, who are evidently less inclined to cite propositions when they are mentioned only once in the discourse. The effect of negotiation for the non-native Finn and Japanese listeners is more obvious. Getting a second hearing in interaction serves to provide extra clues and is the basis for making inferences about words that might have only been partially recognisable in the first hearing (Rost 1990; Rost and Ross 1991).

**Table 5: Negotiation study — covariate analysis**

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum-of-squares</th>
<th>DF</th>
<th>Mean-square</th>
<th>F-ratio</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Form</td>
<td>102.855</td>
<td>1</td>
<td>102.855</td>
<td>35.636</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>NL</td>
<td>106.525</td>
<td>2</td>
<td>53.263</td>
<td>18.454</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Form*NL</td>
<td>2.298</td>
<td>2</td>
<td>1.149</td>
<td>0.398</td>
<td>0.6720</td>
</tr>
<tr>
<td>Covariates</td>
<td>1203.997</td>
<td>1</td>
<td>1203.997</td>
<td>417.148</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>1821.229</td>
<td>631</td>
<td>2.886</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Replication study**

As noted above, the quasi-experimental design incorporated a built-in replication of the ‘source’ negotiation study. The replication simply reversed the order of presentation in the parallel classes sampled so that the baseline version of Dialogue 2 occurs in Form B and the negotiated version in Form A. In the replication study, scores from the primary negotiation study served as a second covariate in the analysis.

The replication surpasses the main effect observed in the negotiated study. The null hypothesis is again rejected in the direction of there being a difference in favour of the negotiated version of the dialogue. Table 6 provides the main effect means for the replication study.

**Table 6: Means and standard error for replication study**

<table>
<thead>
<tr>
<th></th>
<th>Adjusted LS mean</th>
<th>SE</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Form A: Dialogue 2 (Negotiated)</td>
<td>6.094</td>
<td>0.141</td>
<td>337</td>
</tr>
<tr>
<td>Form B: Dialogue 2 (Baseline)</td>
<td>4.538</td>
<td>0.146</td>
<td>301</td>
</tr>
</tbody>
</table>

The main effects in the replication are conditioned by an interaction between native language and form. Although there is ample evidence that negotiated discourse influences comprehension, the interaction of native language and form is different from the ‘source’ negotiation experiment, which found no such interaction. In the replication, it is the Australians and Finns who benefit from the negotiation more dramatically than the Japanese (Table 7). This may be owing to differences in the degree of novelty in the presumably unknown lexical items in Dialogue 2 relative to Dialogue 1, making the negotiated version more salient on the replication study than in the original.

**Table 7: Analysis of variance in replication study**

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum-of-squares</th>
<th>DF</th>
<th>Mean-square</th>
<th>F-ratio</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Form</td>
<td>201.461</td>
<td>1</td>
<td>201.461</td>
<td>65.876</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>NL</td>
<td>575.381</td>
<td>2</td>
<td>287.691</td>
<td>94.072</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Form*NL</td>
<td>37.134</td>
<td>2</td>
<td>18.567</td>
<td>6.071</td>
<td>0.0020</td>
</tr>
<tr>
<td>Covariates</td>
<td>1275.714</td>
<td>1</td>
<td>1275.714</td>
<td>417.148</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>1929.712</td>
<td>631</td>
<td>3.058</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A plausible account for this dependence of native language and impact of negotiation is the salience of the novel lexical content. There may be a ‘floor effect’, mainly for the low proficiency Japanese, that attenuates the impact of negotiated discourse because the novel lexical content remains unprocessible, even with elaboration. The negotiation turn constitutes a ‘push down’ (Varonis and Gass 1982) that may obscure the overall flow of the discourse to these listeners and decrease the likelihood of recall.

**Summary of negotiation study and replication**

The results of the source study and its replication concur in supporting the hypothesis that negotiation of discourse will aid listeners in comprehending more propositions embedded in the stream of speech. The impact of the negotiation of discourse is mediated most strongly by proficiency level. This finding corroborates other studies of discourse negotiation (Rost and Ross 1991; Yano, Long and Ross 1994), which examined the impact of accommodation in listening and reading and the redundancy it engenders. Figure 1 shows the mean differences between the dialogues. Note that the shading indicates difference in the discourse — either baseline or negotiated.
according to the dictation covariate scores. Consequently, the total number of
listeners was reduced to 588 — split between those who had heard Form A or
Form B listening tests.

The focus of the accent study was not, as in the previous study, variation in
group differences in mean comprehension according to the presence of negotia-
tion, but rather, difficulty of citing propositions uttered by either the Australian
or Thai speakers. This difficulty was estimated by splitting the matrix of
dichotomous responses into two independent data sets — subjects hearing
Form A matched with subjects hearing Form B. The two different matrices of
responses were then submitted to Rasch analysis for item difficulty estimation.
The result of these two analyses was a pairing of items derived from the two
forms of the listening test used. It should be noted that since the propositions
appearing on Form A are compared with those on Form B, the effects of
the baseline versus negotiated versions are counterbalanced. Form A was comprised
of a baseline dialogue followed by a negotiated parallel dialogue. For the
matched group hearing Form B the same pattern was followed. The expectation
was, therefore, that the variance attributable to the effect of negotiation will be
eliminated when the two logit estimates from Rasch Analysis were compared
as a repeated measures factor. The remaining fixed effect factor in this design,
a dummy coding for each speaker’s voice in each version of the dialogue,
constitutes the accent factor, which is independent of the effect of negotiation.

Table 8 details the main effects for the repeated measures analysis. The factors
of interest here are the effect of the accent (Australian versus Thai) and the
between-item variation. As can be observed, there is no main effect attributable
to accent differences between the Australian and Thai speakers. There are two
primary interpretations of this fact. The two Thai English as a second language
speakers recorded in the dialogues, though retaining an obvious Thai-English
accent, use ‘standard’ syntax and morphology, thus making the effect of their
accent minimally salient. The other interpretation is that the redundancy pro-
vided in the negotiated discourse serves to compensate for non-hearings in the
baseline version, thus reducing the effects attributable to accent. Essentially,
differences in item difficulty attributable to accent can be smoothed out by
making the dialogue content comparable to authentic accommodated discourse,
thereby giving listeners redundancy in interpreting the accent of the non-native
speaker. This principle, of course, applies to the comprehension of native speak-
er discourse. As Rost and Ross (1991) have shown, queries from listeners allow
a second opportunity to process input that might otherwise go partially decoded.
Table 8: Two-factor repeated measures ANOVA

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum-of-squares</th>
<th>DF</th>
<th>Mean-square</th>
<th>F-test</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accent (A)</td>
<td>2.642</td>
<td>1</td>
<td>2.642</td>
<td>1.402</td>
<td>.2463</td>
</tr>
<tr>
<td>Propositions</td>
<td>52.759</td>
<td>28</td>
<td>1.884</td>
<td></td>
<td></td>
</tr>
<tr>
<td>in citation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Repeated</td>
<td>1.667E-6</td>
<td>1</td>
<td>1.667E-6</td>
<td>3.385E-6</td>
<td>.9985</td>
</tr>
<tr>
<td>measure (B)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AB X citations</td>
<td>13.786</td>
<td>28</td>
<td>0.492</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The within-items (repeated) measure also showed no difference. This was to be expected given the counterbalancing effect of aligning all 30 items contiguously in the analysis.

Finally, we examined the plots of item difficulties in logits. Figure 2 shows the average logits calibrated on the matched pairings of listeners. Embedded, though not graphically obvious in this plot, is the difference between items appearing in baseline versus negotiated mode. As the main effects of the negotiation and replication studies indicated, item difficulty is affected when there is accommodation and negotiation in the discourse. Close examination of the pairs of items would suggest that negotiation can produce sometimes dramatic contrasts in item difficulties, though differences between the accents of the interlocutors in the dialogues apparently did not.

Conclusions

The studies presented here provide clearly replicated evidence of the impact of accommodative negotiation in test discourse. Such negotiation can be manipulated by test designers to control the difficulty of the discourse and accessibility to propositions referred within the discourse. While the issue of redundancy may not be considered important when the content of the listening test is premised on a speech community consisting exclusively of adult native speakers, the present research suggests that when listening tests are devised to simulate authentic interaction in international or multicultural contexts, the monolingual speech community model may well be insufficient. The presence of non-native accents increases both the content and construct validity of listening tests of English for international communication. This conclusion therefore leads us to endorse the inclusion of non-native speakers as participants in the types of interactions from which test content is drawn. Language assessments devised to sample authentic contexts of interaction, tests such as access, can enhance their content validity by including non-native speakers.

Coupled with manipulation of redundancy through accommodation, the comprehensibility of non-native accents does not appear to present an insurmountable problem for test design. Items can be calibrated in terms of their schematic, discoursal, morphosyntactic and phonological features and set down on a common scale characteristic of many item bank schemes. These features could become facets of item attribution as in current applications of ‘rule space methodology’ (Buck and Tatsuoka 1995; 1996). The inclusion of non-native accents may in addition serve to define the content domain of tests purporting to assess English as it is used internationally. Facets of item difficulty may be manipulable by varying the degree of accentedness and the extent of interactional modification embedded in the discourse. The implications of this research suggest that passages containing non-negotiated input and an opaque non-native accent would constitute a very challenging listening task. The same accent could be made more comprehensible by incorporation of some form of recasting of key lexical content into the interactionally modified discourse. Future research on the influence of accent should manipulate varying degrees...
of phonological and morphosyntactic differences from native varieties of English. Concomitant manipulation of the redundancy factors discussed in this paper will aid test designers in devising innovative and content valid test specifications.

References


Appendix 1: Answer sheet for Form A (and B)

Listening comprehension test

Note: READ ALL of the instructions on the test paper. You will also hear instructions on the tape. The tape will be played from start to finish without stopping.

NAME: ____________________________________________________________

DATE: _____________________________________________________________

UNIVERSITY: _______________________________________________________

DEPARTMENT: _______________________________________________________

Part 2

You will now hear a conversation between a man and a woman. They are old friends. As you listen take notes and then summarise the conversation in ENGLISH. Summarise the information about each of the speakers in as much detail as possible.

The recording will be played once only. You have 3 minutes to complete your summary in [L1].

Notes

Write the summary in [L1] about Jim [Kevin] (man) here.

_____________________________________________________________________
_____________________________________________________________________
_____________________________________________________________________
_____________________________________________________________________

Write the summary in [L1] about Cathy [Beth] (woman) here.

_____________________________________________________________________
_____________________________________________________________________
_____________________________________________________________________
_____________________________________________________________________

Part 3
You will now hear a conversation between a man and a woman. They are old friends. As you listen take notes and then summarise the conversation in [L1]. Summarise the information about each of the speakers in as much detail as possible.
The recording will be played once only. You have 3 minutes to complete your summary in [L1].

Notes

Write the summary in [L1] about Kevin [Jim] (man) here.

Write the summary in [L1] about Beth [Cathy] (woman) here.

Part 4
Dictation
You will hear a short dictation of about 40 words. The dictation will be played three times.
The first time you are to listen ONLY.
The second time write down in ENGLISH what you hear as the tape is playing.
The third time check your work.

Write the dictation in ENGLISH here.

You now have one minute to check your work.
The test is over.
Thank you.
Appendix 2

N = Narrator, S = Dictation reader, C = Cathy, J = Jim, B = Beth, K = Kevin, FX = sound effects

Script: Form A


Pause (5 seconds).

N: Part 1. Dictation. You will hear a short dictation of about 30 words. The dictation will be played three times. The first time you are to listen ONLY. The second time write down in English what you hear as the tape is playing. The third time check your work as you listen.

Now listen carefully to the text.

S: (At normal speed) Many people assume that driving a car is a convenience. But when they find themselves in long traffic jams, the apparent convenience of the private automobile is often no longer appealing.

Pause (5 seconds).

N: Now listen once more and write down the text in English as you listen.

S: (Read phrase by phrase) Many people assume that driving a car is a convenience. But when they find themselves in long traffic jams, the apparent convenience of the private automobile is often no longer appealing.

Pause (30 seconds).

N: Now listen once more at normal speed and check your work as you listen.

S: (At normal speed) Many people assume that driving a car is a convenience. But when they find themselves in long traffic jams, the apparent convenience of the private automobile is often no longer appealing.

Pause (15 seconds).

FX: Sound (Beep).

N: That is the end of Part 1. Now turn to Part 2 on page 2.

N: Part 2. You will now hear a conversation between a man and a woman. They are old friends. As you listen summarise the conversation in Finnish. The recording will be played once only.

N: Part 2. You will now hear a conversation between a man and a woman. They are old friends. As you listen summarise the conversation in Japanese. The recording will be played once only.

Pause (10 seconds).

FX: Sound (Beep).

C: Jim?
J: Ohhhh Cathy?, I haven’t seen you for ages. How long has it been?
C: Well, since we finished high school. Has it been ten years?
J: I can’t believe it. You haven’t changed at all!
C: So how have you been? What are you doing?
J: I’m a civil engineer now. I work for the government inspecting public buildings.
C: Really? How did you end up doing that?
J: When I graduated, I got a job diagnosing structural weaknesses in old buildings that might not hold up in a large tremor.
C: Wow. Is it dangerous?
J: Not really, but we have to be careful with viruses. I’m working in the contagious diseases unit now, so we are well prepared.
C: Hey, they are calling my flight. It was nice seeing you again.
J: Nice to see you again, Jim. Take care.
B: You mean you check if the planes are safe? Isn’t that dangerous?
K: No, it’s a desk job actually, I work with computers mainly. What about you? Are you working?
B: I got a degree in nursing and started working at the Children’s Hospital. Then I enrolled in a graduate program specialising in neonatology, ’cause I really like working with newborns.
K: So what do you do exactly?
B: I mostly monitor the incubators according to how well the babies are breathing.
K: That sounds challenging!
B: Yes, it is challenging, especially when we see some very tiny premature babies fight against all odds — when they have no natural immunity against bacteria so we have to keep their environment completely sterile.
K: You mean you isolate the babies from the outside world? Aren’t you afraid you’ll infect them, anyway?
B: It does happen, but we wear gowns and masks all the time which certainly helps keep the germs away from the babies.
K: I think I prefer inspecting airplanes! Well, I’ll have to go now! Beth it was nice seeing you again after all this time.
B: Same here. See you again, I hope, bye Kevin.

J: Bye.
N: Now you have a short time to complete your summary in the space provided.
Pause (3 minutes).
FX: Sound (Beep).
N: That is the end of Part 2. Now turn to Part 3 on page 3.
N: Part 3. You will now hear a conversation between a man and a woman. They are old friends. As you listen summarise the conversation in Finnish. The recording will be played once only.
N: Part 3. You will now hear a conversation between a man and a woman. They are old friends. As you listen summarise the conversation in Japanese. The recording will be played once only.
Pause (10 seconds)
FX: Sound (Beep)
B: Kevin? Is that you?
K: Oh hi Beth. How are you? I haven’t seen you since … our last year of high school.
B: That must be what? … eight years ago.
K: Yes and you haven’t changed a bit!
B: What have you been doing with yourself?
K: Gee where should I begin? I went abroad to study, and now I’m an aviation engineer working for National Airlines inspecting aircraft.
B: Do you test airplanes or something? How did you get into that?
K: Well, I studied physics and I’ve always been interested in aerodynamics, so …
B: What’s aerodynamics?
K: It’s basically the design of aircraft.
B: Oh wow!
K: Yeah, well, it has taken about four years to finish my training — two years to get a degree and another two working as an apprentice aviation engineer. I had to go around helping certified inspectors determine if aircraft were airworthy.

B: You mean you check if the planes are safe? Isn’t that dangerous?
K: No, it’s a desk job actually, I work with computers mainly. What about you? Are you working?
B: I got a degree in nursing and started working at the Children’s Hospital. Then I enrolled in a graduate program specialising in neonatology, ’cause I really like working with newborns.
K: So what do you do exactly?
B: I mostly monitor the incubators according to how well the babies are breathing.
K: That sounds challenging!
B: Yes, it is challenging, especially when we see some very tiny premature babies fight against all odds — when they have no natural immunity against bacteria so we have to keep their environment completely sterile.
K: You mean you isolate the babies from the outside world? Aren’t you afraid you’ll infect them, anyway?
B: It does happen, but we wear gowns and masks all the time which certainly helps keep the germs away from the babies.
K: I think I prefer inspecting airplanes! Well, I’ll have to go now! Beth it was nice seeing you again after all this time.
B: Same here. See you again, I hope, bye Kevin.
N: Now you have a short time to complete your summary in the space provided.
Pause (3 minutes).
FX: Sound (Beep).
N: That is the end of Part 3. Now turn to Part 4 on page 4.
Pause (5 seconds).
N: You will hear a short dictation of about 40 words. The dictation will be played three times. The first time you are to listen ONLY. The second time write down in English what you hear as the tape is playing. The third time check your work.
Now listen carefully to the text.
S: (At normal speed) Thousands of people die each year because they don’t wear seat belts when travelling in automobiles. With the rise in road
fatalities road safety authorities think that there must be an increase in both education and police patrols.

N: Now listen once more and write down the text in English as you listen.
S: (Read phrase by phrase) Many people assume that driving a car is a convenience. But when they find themselves in long traffic jams, the apparent convenience of the private automobile is often no longer appealing.

Pause (30 seconds).

N: Now listen once more at normal speed and check your work as you listen.
S: (At normal speed) Many people assume that driving a car is a convenience. But when they find themselves in long traffic jams, the apparent convenience of the private automobile is often no longer appealing.

Pause (15 seconds).

FX: Sound (Beep).

N: That is the end of Part 1. Now turn to Part 2 on page 2.

S: (At normal speed) Many people assume that driving a car is a convenience. But when they find themselves in long traffic jams, the apparent convenience of the private automobile is often no longer appealing.

Pause (10 seconds).

FX: Sound (Beep).

N: Part 2. You will now hear a conversation between a man and a woman. They are old friends. As you listen summarise the conversation in Finnish. The recording will be played once only.

N: Part 2. You will now hear a conversation between a man and a woman. They are old friends. As you listen summarise the conversation in Japanese. The recording will be played once only.

Pause (10 seconds).

FX: Sound (Beep).

N: That is the end of part one. Now turn to Part 2 on page 2.

S: (At normal speed) Many people assume that driving a car is a convenience.

But when they find themselves in long traffic jams,
The apparent convenience of the private automobile,
Is often no longer appealing.

Pause (30 seconds).

N: Now listen once more at normal speed and check your work as you listen.
S: (At normal speed) Many people assume that driving a car is a convenience. But when they find themselves in long traffic jams, the apparent convenience of the private automobile is often no longer appealing.

Pause (15 seconds).

FX: Sound (Beep).

N: Now listen once more and write down the text in English as you listen.
S: (Read phrase by phrase) Many people assume that driving a car is a convenience. But when they find themselves in long traffic jams, the apparent convenience of the private automobile is often no longer appealing.
C: So how have you been? What are you doing?
J: I’m a civil engineer now. I work for the government inspecting public buildings.
C: You mean checking safety or something? How did you get that sort of job?
J: Well, I studied geophysics in university, and I had a special interest in seismology …
C: What’s seismology?
J: It’s the study of earthquakes.
C: Fascinating!
J: So when I graduated, I got a job diagnosing structural weaknesses in old buildings that might not hold up in a large tremor. My job is to test the foundations of a lot of old buildings around the city and give an analysis of their chances of severe damage in quakes of different magnitudes.
C: You mean you check to see if buildings will be okay in an earthquake? Is it dangerous?
J: No, not really. Most buildings that I inspect are relatively new anyway.

B: That’s fantastic!
K: Yeah, well, it has taken about four years to finish my training — two years to get a degree and another two working as an apprentice aviation engineer.

B: Is it a dangerous job?
K: No, no, it’s a desk job actually, I work with computers mainly. What about you? Are you working?
B: I got a degree in nursing and started working at the Children’s Hospital. Then I enrolled in a graduate program specialising in neonatology.
K: Wow! That sounds challenging!
B: Yes, it is challenging, especially because some very tiny premature babies have to fight against all odds.
K: I think I prefer inspecting airplanes! Well, I’m afraid I’ll have to go now. Beth it was nice seeing you again after all this time.
B: Same here. See you again, I hope, bye Kevin.
K: Bye.

C: So how have you been? What are you doing?
J: I’m a civil engineer now. I work for the government inspecting public buildings.
C: You mean checking safety or something? How did you get that sort of job?
J: Well, I studied geophysics in university, and I had a special interest in seismology …
C: What’s seismology?
J: It’s the study of earthquakes.
C: Fascinating!
J: So when I graduated, I got a job diagnosing structural weaknesses in old buildings that might not hold up in a large tremor. My job is to test the foundations of a lot of old buildings around the city and give an analysis of their chances of severe damage in quakes of different magnitudes.
C: You mean you check to see if buildings will be okay in an earthquake? Is it dangerous?
J: No, not really. Most buildings that I inspect are relatively new anyway.

B: That’s fantastic!
K: Yeah, well, it has taken about four years to finish my training — two years to get a degree and another two working as an apprentice aviation engineer.

B: Is it a dangerous job?
K: No, no, it’s a desk job actually, I work with computers mainly. What about you? Are you working?
B: I got a degree in nursing and started working at the Children’s Hospital. Then I enrolled in a graduate program specialising in neonatology.
K: Wow! That sounds challenging!
B: Yes, it is challenging, especially because some very tiny premature babies have to fight against all odds.
K: I think I prefer inspecting airplanes! Well, I’m afraid I’ll have to go now. Beth it was nice seeing you again after all this time.
B: Same here. See you again, I hope, bye Kevin.
K: Bye.
Test-taker performance on direct and semi-direct versions of the oral interaction module

Kieran O’Loughlin

Introduction

The research which is the subject of this chapter forms part of a larger study investigating the comparability of the live (face-to-face) and semi-direct (tape-mediated) versions of the access oral interaction module.

The two versions of the oral interaction module are necessary for economy and flexibility, since the human and/or technical resources available in each overseas test centre may vary. Given that test takers in overseas test centres may be assigned to either version, it is important that test takers’ results should not be affected by the particular format they undertake. This chapter describes a study of the performance of two individual test takers on both test versions with the aim of examining the interchangeability of the two versions. Data for this study was gathered from the trialling of the access oral interaction module developed in 1994 by the National Languages and Literacy Institute of Australia (NLLIA), Language Testing Research Centre at the University of Melbourne (see O’Loughlin, forthcoming).

The OPI versus SOPI debate

Previous research on the comparability of direct and semi-direct tests of oral proficiency has concentrated primarily on the Oral Proficiency Interview (OPI) — a direct test — and on its more recently developed semi-direct surrogate, the Simulated Oral Proficiency Interview test (SOPI). Both are widely used in a number of countries to assess oral proficiency in a variety of languages. Most of the research work on these two tests has been initiated in the USA by
the American Council on the Teaching of Foreign Languages (ACTFL) under the auspices of the Interagency Language Roundtable (ILR), a body which coordinates government language training agencies. The research undertaken to date has focused mainly on the comparability of test scores obtained from the two formats (see for example Stansfield 1991; Stansfield and Kenyon 1992) and, more recently, on the elicitation tasks used and the language produced by test takers under the two conditions (Shohamy 1994).

On the basis of research carried out at the Centre for Applied Linguistics (CAL) in Washington DC, Stansfield (1991) suggests that the SOPI has shown itself to be a valid and reliable substitute for the OPI. He reports Pearson correlations of between 0.89 and 0.95 between the direct and semi-direct versions used in various languages. In a later study, Stansfield and Kenyon (1992) used Generalisability Theory to further explore this issue. Generally, low levels of interaction between candidate and test format were found in a range of languages for candidates who had undertaken the two kinds of test. On the basis of their research using test scores, Stansfield and Kenyon (1992) conclude that the OPI and SOPI are highly comparable as measures of oral language proficiency. They assert that they may be viewed as parallel tests delivered in two different formats.

Shohamy (1994) adopts a more sceptical position in relation to statistical comparisons of scores on these two tests. Unlike Stansfield (1991), she argues that high correlations between scores on different kinds of tests provide necessary but insufficient evidence for test substitution and that they may not necessarily be measuring the same trait. She emphasises the need to examine the validity of tests from multiple perspectives, not just from the point of view of test scores, to obtain a deeper understanding of what they actually measure. This provides the rationale for her study of the elicitation tasks used and, in particular, the language produced in the Hebrew OPI and SOPI.

In her discourse-analytic study of language output in the OPI and SOPI, Shohamy found that the SOPI elicited a more limited range of language functions than the OPI and that SOPI answers included more self-correction, repetition of phrases in the eliciting questions, paraphrasing and a more restricted range of prosodic features. The discourse produced in the SOPI was also more formal, cohesive and of a higher lexical density and thus more ‘literate’ than in the OPI.

Shohamy’s interpretation of the various differences which emerged between the OPI and the SOPI is that the test context or format (i.e. whether face-to-face or tape-mediated) rather than the elicitation tasks, is the most powerful determinant of the type of language produced. The general implication is that the OPI and SOPI do not appear to measure the same trait and are therefore not interchangeable as tests of oral proficiency, even where a high correlation exists between scores on the two tests.

The access: oral interaction module

Test design

The access: oral interaction module aims to assess general oral proficiency. The live version is designed to be administered to individual candidates by a trained interviewer in a face-to-face context while the tape version is designed to be delivered to a group of candidates in a language laboratory. Both formats include an unassessed warm up phase. The live version is designed to last 20 minutes and the tape version 30 minutes when administered overseas, although the trial versions are often up to ten minutes longer for the purpose of trialling a greater number of tasks. Test booklets are produced for the interviewer and candidate on the live version and a candidate booklet, tapescript and sound recording for the tape version. A series of matching tasks were developed so that important characteristics, such as task structure and expected range of language functions, should be parallel across the two formats. In both instances the amount of preparation time for each task is set in advance; however, only on the tape version is the amount of response time fixed beforehand.

Test tasks and scoring criteria

Table 1 provides a summary of the tasks, together with the scoring criteria which were used to rate each of them, as they were defined in June 1994 when the trial which is the subject of this study took place. In both test formats these criteria focused on fluency and accuracy. The only difference was the inclusion on the live version of an additional category, comprehension, since the ability to understand the interlocutor was considered more crucial to the candidate’s performance than on the tape version. Whereas on the live version most of the instructions were spoken only, on the tape version the instructions were both spoken and written out in full in the candidate’s booklet. This additional assistance was given on the tape version since candidates would not have the opportunity to clarify the instructions with an interlocutor as they could in the live version.
Each of the scoring criteria were assessed on a seven-point scale with descriptors (these are identical for the two versions of the test). The full list of scoring criteria and their accompanying descriptors are included in Appendix 1.

**Table 1: Test tasks and scoring criteria for access: oral interaction modules (June 1994)**

<table>
<thead>
<tr>
<th>Live Description</th>
<th>Tape Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>fluency</td>
<td>fluency</td>
</tr>
<tr>
<td>grammar</td>
<td>grammar</td>
</tr>
<tr>
<td>vocabulary</td>
<td>vocabulary</td>
</tr>
<tr>
<td>coherence and cohesion</td>
<td>coherence and cohesion</td>
</tr>
</tbody>
</table>

**Narration**
- grammar
- vocabulary
- coherence and cohesion

**Role play**
- Telephone message
- fluency
- grammar
- intelligibility

**Exposition**
- fluency
- vocabulary
- coherence and cohesion

**Discussion**
<table>
<thead>
<tr>
<th>Live Description</th>
<th>Tape Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>fluency</td>
<td>fluency</td>
</tr>
<tr>
<td>grammar</td>
<td>grammar</td>
</tr>
<tr>
<td>intelligibility</td>
<td>intelligibility</td>
</tr>
</tbody>
</table>

**Global criteria**
- communicative effectiveness
- comprehension

**Table 2: Proficiency levels: oral interaction module**

<table>
<thead>
<tr>
<th>Level 6: Vocational+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Can communicate effectively in spoken English in a range of social, educational and work situations. Communication is appropriate with a high degree of fluency. Language is grammatically accurate most of the time with a wide range of vocabulary which is used effectively in most situations.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level 5: Vocational</th>
</tr>
</thead>
<tbody>
<tr>
<td>Can communicate in spoken English to handle everyday communication in social, educational and work situations. Can communicate with a fair degree of fluency despite some grammatical inaccuracies. Vocabulary is wide enough to express most ideas, particularly in familiar situations.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level 4: Functional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Can communicate in spoken English to handle everyday communication in social, educational and work situations. Can communicate with a fair degree of fluency despite some grammatical inaccuracies. Vocabulary is wide enough to express most ideas, particularly in familiar situations.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level 3: Basic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Can communicate general information in spoken English in most everyday social situations. Can use basic grammatical structures although inaccuracies are frequent. Although vocabulary is limited at this level most common concepts can be expressed.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level 2: Limited</th>
</tr>
</thead>
<tbody>
<tr>
<td>Can communicate in spoken English well enough to hold a very simple conversation. Limited control of basic grammatical structures. Vocabulary is limited to common words and phrases.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>No practical speaking ability in English.</td>
</tr>
</tbody>
</table>

**Proficiency levels**

Six general levels of performance on the oral interaction module were distinguished for the purpose of candidate classification, although the highest level was not described. These levels and their descriptors (where applicable) are shown in Table 2. The levels for each individual scoring category (ie fluency, vocabulary, grammar etc.) directly corresponded to these six overall access: performance levels. In each case levels 6 and 7, for each scoring criterion, corresponds to Vocational+, level 5 to Vocational, level 4 to Functional, level 3 to Basic, level 2 to Limited and level 1 to below Limited.

Setting the levels

In terms of reliability of measurement, the test was designed to discriminate most clearly between the three upper levels — Level 4: Functional, Level 5: Vocational and Level 6: Vocational+. The cut-off points for these three levels were derived from a trial in September 1993 where the entire cohort of test takers
Both versions of the access: oral interaction module each test taker’s audio-recorded performance is assessed independently in Australia by two trained assessors. The fact that the rating is done retrospectively using audio-tapes in both cases means that assessment can be carried out in the same way. Typically this is done over a weekend in language laboratories where raters assess a total of about twenty tests (ten each day), some live and some semi-direct.

An integral part of the test development process is the periodic trialling of new forms of the four modules (reading, writing, listening and speaking). An important advantage of these trials for the purpose of this study, is that, unlike the overseas administrations, test takers undertake both versions of the oral interaction module. This allows direct comparisons to be made between performance under the two conditions.

Research into the comparability of the direct and semi-direct versions

Several recent studies have examined the comparability of the direct and semi-direct versions of the access: oral interaction module from various perspectives. Wigglesworth and O’Loughlin (1993), for example, undertook a preliminary investigation into test scores using the multi-faceted Rasch program, FACETS (Linacre 1989–1995). This study was based on data obtained from the first trialling of the test in December 1992 (N= 83) which was designed to identify candidates at the vocational level only. In this trial the live and tape performances of all candidates were independently rated on six occasions to maximise the reliability of measurement, whereas in subsequent trial and overseas administrations each performance was rated only twice. Wigglesworth and O’Loughlin found that, although the candidate ability estimates obtained using the FACETS program were strongly correlated (r= 0.92), ten of the eighty-three candidates (12%) achieved vocational level on one version but not on the other, again using ASLPR ratings to establish the cut-off scores for the two formats. This finding suggested the existence of a significant test method effect.

Hill (see Chapter 6) investigated test-taker feedback and reported that, when asked to compare the two formats, 90 per cent of candidates in the December 1992 trial felt that the live version was the better test of their ability to speak English. Furthermore, 85 per cent of them found the tape test more difficult and 60 per cent felt more nervous on it. These figures strongly suggested that the direct version appealed more to test takers. However, analysis of the other questions focusing on each version separately indicated that the strength of the preference for the live format may not have been as great as the first set of results suggested. On the whole, candidates appeared to be only slightly less...
positive about the tape version overall on a range of categories, including perceived difficulty, fairness, adequacy of preparation and response time, clarity of instructions and whether each of the formats provided a good test of their ability to speak English. Whether or not ‘test appeal’ can be considered an aspect of validity, it is clearly an important practical consideration in making tests acceptable to test takers and test users. This is particularly the case, as Hill suggests in Chapter 6, in tests such as access: where overseas test takers have a large investment in the outcome.

O’Loughlin (1995) examined candidate output under the two conditions from the perspective of lexical density — a measure of the relationship between lexis and grammar in spoken or written discourse. The study focused on the same trial as Wigglesworth and O’Loughlin (1993) and Hill (Chapter 6), in this case using the audio-recorded test performances as data. The investigation suggested that the live version tapped language of a significantly lower (in statistical terms) level of lexical density than the tape version, although the overall difference was not as great as in Shohamy’s (1994) study of the Hebrew SOPI and OPI. O’Loughlin (1995) concluded that the degree of interactiveness was probably the major determinant of lexical density in candidate output on the two versions. It was also suggested that there are probably a number of factors — including the nature and quality of interlocutor talk, task structure, and preparation and response time — which influence candidate output in oral proficiency tests, not simply whether the candidate is talking to another person face-to-face or to a microphone per se. Finally, given the strength of the correlation between the candidate ability estimates obtained from each version (ie r=0.92), it was tentatively suggested that the two versions were still likely to be measuring the same language trait despite the observed differences in the discourse produced by candidates under the two conditions.

**Rationale for the study**

Language testing research has traditionally focused on group performance in *a posteriori* investigations of validity and reliability. This study, however, examines how the investigation of individual test takers may assist in addressing a major concern in measurement (identified by Bachman 1990) — that of identifying, estimating and controlling the effects of factors that affect test scores. This issue, Bachman explains (1990: 160–161), relates to both reliability and validity:

> The investigation of reliability is concerned with answering the question, ‘How much of an individual’s test performance is due to measurement error, or to factors other than the language ability we want to measure?’ and with minimising the effects of these factors on test scores. Validity, on the other hand, is concerned with the question, ‘How much of an individual’s test performance is due to the language abilities we want to measure?’ and with maximising the effects of these abilities on test scores.

The factors ‘other than the language ability we want to measure’ include:

- test method facets (such as whether the input is live or taped in the case of a speaking test or the different topics of individual tasks);
- random factors (such as the test taker’s mental state or changes in the time of day that a test is carried out);
- attributes of the test taker not considered to be related to language ability (such as sex, ethnic background, cognitive style).

The first two of these are considered to be possible sources of measurement error and the third to be a source of test bias or invalidity (Bachman 1990: 166) Other systematic factors affecting test scores include the consistent differences in behaviour between different interlocutors on direct tests of oral proficiency, and between different raters where test scores are obtained subjectively — such as in direct assessments of speaking and writing. Shifting the focus of study to individual test takers has the potential for providing deeper insights into how these factors impact on test scores and into what language tests actually measure.

**Method**

**Overview**

This investigation employs a case study approach to explore the issue of test comparability. Like ethnography, this approach is based on the assumption that human behaviour, in this instance test performance, cannot be understood without incorporating into the research the perceptions and insights of those involved in the study, and not simply those of the researcher. Like ethnography, it employs observation, interviews and other sources of qualitative data as methods of data collection. Case studies, however, are generally more limited in scope than ethnographies, they focus less on cultural interpretation and more frequently use both qualitative and quantitative data and analyses (Nunan, 1992). The study focuses particularly on two trial test takers who undertook both live and tape-mediated versions of the access: oral interaction module.
The trial population

The various types of data used in this study were gathered from a trial held on 4 June 1994 in Melbourne in which 94 volunteer candidates recruited from a number of local English language centres and tertiary institutions completed all four modules of a new form of the *access*: test. Each test taker was paid AUD$70 upon completion of the trial test.

The trial test taker population represented diverse cultural and linguistic backgrounds as well as a wide range of age groups. It is important to note that none of the test takers in this and other trials exactly fit the description of the target group for whom the *access*: test was designed; that is, qualified, usually experienced, professionals of non-English speaking background (NESB) who are over twenty-five years of age and are seeking permanent residence in Australia from outside the country. The most substantially ‘non-target-like’ group who take part in the trials is normally NESB students, under twenty-five and living in Australia on a temporary basis while completing their tertiary studies. This group constituted 58 per cent of the test cohort in the June 1994 trial. The remaining 42 per cent belonged to the ‘target-like’ group which was composed of NESB professionals qualified in a wide range of fields including engineering, medicine, teaching, administration, computing, librarianship and architecture. However, in so far as individuals in this group had mostly already obtained permanent residence, had been in Australia for more than 18 months and were undertaking the test in Australia, they did not exactly represent the target test-taker population either.

The numbers for the two groups of candidates in the June 1994 trial contrast with the December 1992 trial studied by Wigglesworth and O’Loughlin (1993) where the figures for the non-target-like and target-like groups were 32 per cent and 68 per cent respectively. In retrospect, it would have been preferable if the test cohort in the June 1994 trial had more closely resembled the one in the earlier trial so that data obtained from the trial, including test scores and test taker feedback, could have been used to make more meaningful predictions about the performance of overseas candidates on this particular form of the test. Given these differences between the trial and overseas test populations and the fact that trial candidates obviously do not have the same investment in the outcome of the test as those people undertaking it in the real test situation overseas, any conclusions reached on the basis of the findings in this study can only tentatively be transferred to the actual test situation.

Data collection

The study involved close observation of two candidates’ performance under both the live and tape-mediated conditions. Subsequent interviews with each of them, with their respective interlocutors (in the live version) and with the six raters who later assessed them were also included as methods of data collection. This triangulation provided a series of ‘snapshots’ of the two candidates taken from different perspectives. Together, these pictures yielded rich insight into the relationship between the two test takers’ performances and their scores on both versions. Responses to questionnaires provided by 77 of the 94 test takers formed additional data about test taker reactions to the two tests. The test scores of the candidates were then examined in relation to the rest of the test cohort. Further data collection and analyses were later undertaken to check the reliability of the original ratings and to gain deeper insight into some of the factors affecting the test scores of the two candidates in each version. This process involved additional assessments of the two candidates under both test formats and detailed comments given by an extra rater. Other more sophisticated statistical investigations of test score data were also undertaken.

Subject selection

In order to identify two suitable people for the study, six test takers were approached before they undertook both versions of the oral interaction test on the day of the trial administration. They were informed of the purpose of the study and the nature of their participation; that is, that they would be observed while completing the two versions and then interviewed afterwards. They were also made aware that their test performances would be videotaped and their interviews audiotaped, but that confidentiality with respect to their identities would be maintained at all stages of the project. Given the impossibility of representativeness in studying only two test takers out of a large test cohort, the only criteria for selection were that one of these people would have done the live test first and the other the tape test and that one would be male and the other female.

Two very different people, henceforth referred to by the pseudonyms ‘Abdul’ and ‘Wing Li’ ultimately agreed take part in the project. Abdul was an Egyptian male, fifty-two years of age who had worked as a TV director in Egypt before arriving in Australia and taking up permanent residence in 1992. He was unemployed at the time of the trial and studying general English full-time. Wing Li
was a Chinese female, twenty-five years of age, who arrived in Australia as a temporary resident in 1993. She was a student of tourism and hospitality at the time of the trial. Both candidates reported having studied English for ten years before coming to Australia. Abdul completed the live then tape version of the speaking module in the morning before undertaking the rest of the test. By contrast, Wing Li did the tape then live version of the speaking module in the early afternoon after undertaking the reading and writing modules, but prior to completing the listening module. Abdul was interviewed by a male ESL teacher in the live version and Wing Li by a female, non-ESL trained primary school teacher.

**Findings**

**Observation**

Abdul appeared to perform better in the live than in the tape version at least partly because of the role played by his interlocutor in the interaction. The interlocutor displayed much patience (allowing the candidate plenty of time to respond) and gave reassurance and positive feedback of both a verbal and non-verbal nature. On the other hand, Wing Li seemed to perform less well on the live version, also at least partly because of her interlocutor who tended to rush through the test and provide minimal verbal and non-verbal feedback to the candidate. The success of the interaction therefore clearly has important consequences for the candidate’s overall performance.

The quality of the input may be the same on the tape version, but it seems there are other variables impacting significantly on the test taker’s performance. Wing Li appeared to adapt much better to the more constrained conditions of the tape test than Abdul who often appeared unable on nearly all of the tasks to process the requirements of the tasks and then produce a response in the time allowed. Both candidates were frequently cut off by the next question, but Wing Li usually seemed closer to completing her answer. Abdul, on the other hand, often seemed to be still gathering his thoughts when the signal to begin talking was given and only at the mid-point of his response when stopped short by the recorded voice. It should be noted that Wing Li may have been better prepared for the tape version of the speaking sub-test than Abdul by virtue of the fact that, unlike him, she had already completed the listening module conducted in the language laboratory earlier in the day. Another possibility raised by Hill (Chapter 6) is that candidates from Asian countries (such as Wing Li) may be more familiar with language laboratories which are widely used in language teaching.

**Candidate interviews**

Interviews were conducted immediately after each candidate had completed the two versions of the oral interaction test. Abdul reported feeling more nervous on the tape test. He also reported finding the tape test more difficult. He felt it was easier to talk to another person than to a microphone because clarification questions could be asked (this he did on several occasions) and there was time to respond. He was very satisfied with his interlocutor who built his confidence by giving him ample time to process and respond on each task. On the tape version he reported needing more time to prepare for some questions, feeling frustrated at being unable to ask clarification questions and also pressured to finish his answers before being cut off by the next task. In general, therefore, he felt the live version was a better test of his ability to speak English.

Wing Li preferred the live version overall. She felt more confident on the tape version but liked the fact that there was an interlocutor who responded in the live version. She also felt that the live version gave her more opportunity to talk. On the other hand, she found the live tasks slightly harder. She was more nervous in the interview as she was anxious about whether she would understand the interlocutor and, more particularly, whether the interlocutor would comprehend her. Thus, in the live test, she felt more pressure to monitor her choice of syntax and vocabulary to ensure she was understood properly. This was not the case in the tape version. In general, therefore, she felt the live version was a much better test of her proficiency because she was more aware of the need to communicate and therefore tailor her language appropriately. Surprisingly, she expressed no dissatisfaction about her interviewer’s performance, a view which was at odds with my own opinion.

**Candidate questionnaires**

Of the ninety-four candidates who undertook both versions, seventy-seven completed a questionnaire on the oral interaction test. Table 3 is a summary of their reactions to the main questions, together with the matching figures, where applicable, from the earlier study undertaken by Hill (Chapter 6) of the December 1992 trial. There was a clear preference for the live test in both trials. However, despite the fact that the overwhelming majority of candidates thought the live version a better test, it is worth noting that more than half of those in the 1994 trial still felt the tape version was a good test of their spoken English.
Both candidates felt the live version gave them more opportunities to speak English and that it was a better test of their spoken English. Abdul felt more nervous on the tape test and also found it more difficult. Wing Li, on the other hand, felt more nervous in the live version and found it more difficult than the tape.

In their responses to other more detailed questions dealing with the tape version separately, both candidates reported that the tape was always audible, the instructions were clear, that the test was pitched at the appropriate level of difficulty overall but that there was sometimes insufficient time given to complete their answers. Abdul did not always have enough time both to prepare and to answer the questions and was not sure whether it was a good test of his spoken English. Wing Li, on the other hand, felt there was always adequate preparation time but sometimes insufficient time to respond on the tape. She also thought that the tape test was not a good test of her spoken English.

The responses of the two candidates to the parallel questions on the live version were identical. They both always felt they understood the instructions and had enough time to prepare and answer the questions. They thought the interlocutor was consistently helpful, that the test was pitched at the appropriate level of difficulty overall and that this version was a good test of their spoken English.

Interlocutor interviews
When questioned about his perceptions of the candidate’s performance in the live test, Interviewer 1 thought that Abdul did well and improved as the test progressed. The interviewer felt that he performed well as an interlocutor, that he had not been too intrusive and had allowed the candidate sufficient time to complete his responses. In this instance his self-assessment was consistent with my own observations about his performance. He claimed not to be distracted by the presence of an observer and the video camera.

Interviewer 2 thought that Wing Li performed fairly well but that she seemed more comfortable with the last two sections of the test, the exposition and discussion tasks. She also felt that her own performance as an interlocutor was creditable, that she had been encouraging and had allowed the candidate the opportunity to respond fully on each of the tasks. This self-assessment ran contrary to my own evaluation of her performance. Although at the time of the interview she suggested that the video camera had not distracted her unduly, she later reflected that it had made her uneasy and that this may have negatively affected her interaction with the candidate.

---

### Table 3: Candidate reactions — June 1994 and December 1992 Trials

<table>
<thead>
<tr>
<th></th>
<th>June ’94</th>
<th>Dec ’92</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Which test made you feel more nervous?</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tape</td>
<td>68%</td>
<td>60%</td>
</tr>
<tr>
<td>Live</td>
<td>27%</td>
<td>40%</td>
</tr>
<tr>
<td>Same</td>
<td>5%</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>June ’94</th>
<th>Dec ’92</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Which test was more difficult for you?</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tape</td>
<td>95%</td>
<td>85%</td>
</tr>
<tr>
<td>Live</td>
<td>5%</td>
<td>15%</td>
</tr>
<tr>
<td>Same</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>June ’94</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Which test gave you more opportunity to speak English?</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tape</td>
<td>5%</td>
<td></td>
</tr>
<tr>
<td>Live</td>
<td>95%</td>
<td></td>
</tr>
<tr>
<td>Same</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>June ’94</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Do you think the tape version was a good test of your spoken English?</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>58%</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>26%</td>
<td></td>
</tr>
<tr>
<td>Not sure</td>
<td>16%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>June ’94</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Do you think the live version was a good test of your spoken English?</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>94%</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Not sure</td>
<td>6%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>June ’94</th>
<th>Dec ’92</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Which one was a better test of your spoken English?</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tape</td>
<td>94%</td>
<td>90%</td>
</tr>
<tr>
<td>Live</td>
<td>5%</td>
<td>10%</td>
</tr>
<tr>
<td>Same</td>
<td>1%</td>
<td>0</td>
</tr>
</tbody>
</table>
Rater interviews

Six of the twenty-two raters engaged to mark the audio-recordings from this trial were involved in assessing Abdul and Wing Li. The particular raters assigned in the design devised by the test project team to Abdul and Wing Li were as follows:

<table>
<thead>
<tr>
<th>Candidate</th>
<th>Live: Raters</th>
<th>Tape: Raters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abdul</td>
<td>7 and 89</td>
<td>3 and 46</td>
</tr>
<tr>
<td>Wing Li</td>
<td>3 and 72</td>
<td>7 and 39</td>
</tr>
</tbody>
</table>

In accordance with the rating design, none of these raters assessed either test taker under both live and tape conditions. Each rater was subsequently interviewed about the performance of their assigned candidate. While the whole group of twenty-two raters were informed beforehand that they might be interviewed, these six raters were not aware that they would be questioned about either Abdul or Wing Li until they had completed their assessment of them.

The two raters who assessed Abdul’s live test felt he had performed reasonably well and that he had improved as the test progressed. They agreed that his interviewer had been very supportive and effective. Those who rated Abdul’s tape performance both indicated that he seemed ill at ease and would probably have performed better on a live test where he could have been prompted to speak at greater length.

The raters who scored Wing Li’s live test felt that she had performed inconsistently across the various tasks. Interestingly, while one rater applauded the fact that the interlocutor had adhered very strictly to her designated script, the other felt she should have been more flexible and drawn out the candidate more. This suggests that raters may have conflicting views about what is a ‘good interlocutor’ in this test. This is confirmed by Morton and Wigglesworth (1994) in a study where raters were required to classify interlocutor behaviour on the live version of the access: oral interaction module as poor, average or good. They found it was quite common for two raters to disagree on their classifications of an interlocutor’s performance with any particular candidate. The two raters also disagreed about whether Wing Li had provided sufficient output in order to be reliably assessed.

Both raters who assessed Wing Li’s tape test felt the candidate had performed very well overall and that she had seemed quite relaxed and confident in tackling the test. One of the raters reported experiencing difficulty with this candidate distinguishing between levels 5 and 6 on the rating for most of the criteria.

Test scores

A total of 94 candidates completed both versions of the trial test. When trialling new forms of the live and tape versions, several previously used tasks are included as ‘anchors’. This enables candidate performance to be compared with earlier forms of the test. Most importantly, it enables cut-off points for the different overall levels of performance (see Table 2 in Chapter 6) to be carried over from one trial to the next. In this trial the ‘anchors’ were the description and exposition tasks on the live version and the summary task on the tape version.

Multi-faceted Rasch analyses using the program FACETS (Linacre 1989–1995) were run on the data obtained during the access: trial. In this analysis each candidate receives an ability measure known as a logit which expresses the probability of the candidate obtaining a particular score (or above) given the ability of the candidate, the difficulty of the item and the harshness of the rater. For the 1994 trial analysis the item difficulty estimates for each of the criteria on the three previously trialled tasks were anchored; that is, they were fixed into the analysis in advance. Initially, the live and tape data were treated as coming from the same test and combined to obtain ability estimates for each of the 94 candidates. For the purpose of this study, separate analyses of the live and tape data were then carried out, this time anchoring all of the item ability estimates obtained from the combined analysis. This was done so that the established cut-off scores would apply to the results from the two formats taken separately.

As in Wigglesworth and O’Loughlin’s (1993) study of the first trial held in December 1992, the number of ‘misfitting’ candidates (those whose scores did not conform to the overall pattern of scoring for the test cohort overall and whose abilities were therefore not being measured appropriately by the test) was relatively small — on average about 3 per cent for the three previously trialled tasks were anchored; that is, they were fixed into the analysis in advance. Initially, the live and tape data were treated as coming from the same test and combined to obtain ability estimates for each of the 94 candidates. For the purpose of this study, separate analyses of the live and tape data were then carried out, this time anchoring all of the item ability estimates obtained from the combined analysis. This was done so that the established cut-off scores would apply to the results from the two formats taken separately.

The candidate ability estimates obtained using the live and tape data taken separately correlated only moderately well at r = 0.80. This contrasts with the higher figure (r = 0.92) obtained in Wigglesworth and O’Loughlin’s (1993) study. Possible reasons for this discrepancy will be discussed in the conclusion.
on test scores given that candidates undertake both formats on the same day. (It should be remembered at this point that candidates in overseas administration only complete one version.) Thirty-three of the seventy-seven candidates in this trial who completed the questionnaire undertook the live version first and forty-four undertook the tape version. Forty-four per cent scored higher on the version they did first, a result which suggests that there was not a significant practice effect overall (in either a positive or negative sense) on candidate performance across the two versions.

The cut-off logit scores for the three key levels of proficiency targeted by the test were as follows:

- Level 4: Functional = -0.8
- Level 5: Vocational = 0.8
- Level 6: Vocational+ = 1.8.

Candidates were ultimately classified according to the level of proficiency they attained in both versions on the basis of their logit scores in each case. Table 4 summarises this information. The results indicated that 40 per cent of all candidates were assigned a different level of proficiency on the two versions of the test with 28 per cent doing better on the tape test and 12 per cent on the live test. In some cases (11 per cent of all candidates) there was a discrepancy of more than one level. These results indicated relatively poor consistency of measurement across the two formats in this instance. Again, these findings contrast fairly unfavourably with Wigglesworth and O’Loughlin’s (1993) study of the 1992 trial where 12 per cent of candidates were allocated a different proficiency level. Possible reasons for the poorer level of correspondence between the results from the two versions of the test in this trial compared to the earlier one are discussed in the conclusion.

In the cases of Abdul and Wing Li there is a sharp discrepancy between their live and tape ability estimates as shown in Table 5. However, in Abdul’s case the difference did not affect the proficiency level to which he was assigned in the final analysis.

The discrepancy between Wing Li’s logit scores on the two formats is clearly the more disconcerting set of results as the access: oral interaction module aims to reliably identify candidates at the Functional, Vocational and Vocational+ levels. It was not designed to distinguish accurately between candidates below functional level. While the difference in Wing Li’s results turns out to be highly
significant, both of Abdul’s results can simply be read as his not having achieved functional level on either version.

Table 4: Correspondence between candidate proficiency levels on the live and tape versions of the oral interaction module, June 1994 trial (N=94)

<table>
<thead>
<tr>
<th></th>
<th>Tape</th>
<th>Live</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-F</td>
<td>F</td>
</tr>
<tr>
<td>-F</td>
<td>24</td>
<td>14</td>
</tr>
<tr>
<td>Live</td>
<td>F</td>
<td>3</td>
</tr>
<tr>
<td>V</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>V+</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

- F = below Functional
F = Functional
V = Vocational
V+ = Vocational+

Table 5: Proficiency levels achieved by Abdul and Wing Li on live and tape formats

<table>
<thead>
<tr>
<th>Candidate</th>
<th>Test version</th>
<th>Logit</th>
<th>Proficiency level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wing Li</td>
<td>tape</td>
<td>1.84</td>
<td>Vocational+</td>
</tr>
<tr>
<td></td>
<td>live</td>
<td>-1.22</td>
<td>below Functional</td>
</tr>
<tr>
<td>Abdul</td>
<td>live</td>
<td>-1.56</td>
<td>below Functional</td>
</tr>
<tr>
<td></td>
<td>tape</td>
<td>-2.64</td>
<td>below Functional</td>
</tr>
</tbody>
</table>

The additional rater

As noted earlier, the rating design for the trials was such that no rater assesses any individual test taker twice; that is, under both live and tape conditions. One of the most experienced and reliable raters (Rater 12), who had not been assigned either of the candidates during the rating procedure, was therefore asked, several weeks after the other ratings had been completed, to assess both candidates under the live and tape conditions using the video-recordings. After viewing the video-recordings the rater was not only questioned about the test takers’ performances but also invited to comment on the feedback from the candidates and their respective interlocutors and raters. This provided a test of the reliability of my own interpretations of the findings in the study.

Rater 12 ranked the four performances in the same order as the ability estimates derived from the FACETS analysis indicated. However, his total raw scores did not suggest as wide a discrepancy between the live and tape performances of Wing Li in particular. This indicated one or more of Wing Li’s original raters may have been uncharacteristically harsh or lenient in their scoring. This hypothesis was investigated empirically using an extension of the FACETS program known as bias analysis.

Bias analysis identifies unexpected patterns of behaviour which may occur from an interaction of a particular rater with respect to some component of the rating situation such as a particular candidate, format or item. In this case the focus was on the interaction between rater and candidate. A bias analysis which incorporated Rater 12’s raw scores for Abdul and Wing Li into the original data set was then undertaken. The results indicated that two of Wing Li’s sets of ratings were significantly biased and thus unreliable. Rater 39 assessed Wing Li’s tape performance significantly more leniently than for other candidates, while Rater 72 scored Wing Li’s live performance significantly more harshly than for other candidates. These findings would seem to explain, at least in part, the strength of the discrepancy between the live and tape logit scores for this candidate as the ability estimates calculated by the FACETS program do not take this kind of interaction into account. None of the raters who assessed Abdul appeared to be significantly biased in their ratings.

Possible reasons for the erratic behaviour of two of Wing Li’s raters were revealed in the rater interviews. Rater 39 reported difficulty in distinguishing between levels 5 and 6 for many of the criteria and seems to have given the candidate the benefit of the doubt on about half the categories. This may explain why his ratings for Wing Li emerged as significantly more lenient than was usual for him. In Rater 72’s case, she seems to have heavily penalised Wing Li for insufficient output linked to what she perceived to be the poor interactive competence of both the candidate and her interlocutor. Rater 3, who also assessed Wing Li’s live performance, expressed no such criticisms of either the candidate or interlocutor.

The additional assessor, Rater 12, provided further valuable insights into the performances of Wing Li. He suggested that her poorer performance on the live version might be explained both by her discomfort in undertaking a test involving interaction and by not being well supported by her interlocutor. Because the candidate was nervous herself she may not have been aware of how much...
the interlocutor was feeding into her anxiety by rushing through the test and providing minimal support in the interaction. In any case, the interlocutor’s more formal, less supportive demeanour may have been in accordance with the candidate’s expectations about the behaviour of examiners in oral language tests. These reasons might then explain why she expressed no dissatisfaction with her interlocutor afterwards. On the tape version, however, Wing Li seemed more confident and relaxed and therefore more able to demonstrate her proficiency. Rater 12 agreed that Abdul’s interlocutor more successfully drew him out with encouragement, patience and positive feedback. This candidate seemed less comfortable on the tape version although his overall performance was judged to be only marginally better on the live version.

Conclusion

The finding that 40 per cent of all candidates were assigned a different overall level of proficiency on the direct and semi-direct versions of the access: oral interaction module in the June 1994 trial suggests that the measurement process was not sufficiently constrained so as to yield a satisfactory level of classification consistency across the two formats. Of course, it may be that the results from the study of the December 1992 trial by Wigglesworth and O’Loughlin (1993), which suggested a somewhat stronger correspondence between the test scores on the two versions, represent a more accurate measure of this relationship given that the trial test cohort more closely resembled the target overseas test population and that the measurement error associated with inconsistent rater behaviour was substantially reduced by increasing the number of independent ratings from two (as is normally the case) to six for each candidate. Yet, despite the stronger correlation \( r = 0.92 \), Wigglesworth and O’Loughlin (1993) found that 12 per cent of candidates did not reach the same overall proficiency level under the two test conditions. This finding suggested a significant test method effect, especially given that in the earlier trial, unlike later trials and overseas administrations, candidates were assigned to either one of only two proficiency levels — Vocational and below Vocational. The results from the current study could therefore be interpreted as confirming the lack of equivalence between scores from the two versions suggested in the earlier study.

However, there are several reasons why the relatively low levels of correspondence between the test scores on the two versions found in both these studies should be interpreted with caution. Firstly, the fact that a substantial proportion of candidates were classified differently by the two versions of the oral interaction module in both trials may have been partially a function of the way in which the cut-off scores were established. Standard setting in criterion-referenced test design has been identified as a major ongoing source of concern by a number of measurement specialists (see, for example, Cizek 1993) and there is ongoing debate concerning the appropriacy and utility of the various methods which can be employed (Berk 1995). Further, research in this area indicates that different judgemental standard-setting methods lead to different classifications (Berk 1996: 216). It may be, therefore, that the degree of correspondence between results on the two versions of the test would have been higher had a different method of standard setting been used. Secondly, the transferability of the findings in the two studies to the overseas test context may be jeopardised to some degree by differences between the composition of the trial and overseas test cohorts, particularly in the trial on which this current study was based. In addition, it is worth remembering that trial candidates, unlike overseas test takers, undertake both versions of the test and obviously do not have the same investment in the outcome of the test as those people undertaking the test overseas. Taken together, these considerations suggest that any conclusion concerning the comparability of the direct and semi-direct versions in the real test situation based on the findings from studies of trials can only be tentatively drawn.

Notwithstanding these caveats on the overall findings in this study, the observational, interview, the questionnaire and the test score analyses of individual candidates suggest that there are a variety of factors in the testing process currently contributing towards measurement error on the two formats. The most salient of these factors appear to be:

- the candidate’s level of familiarity and comfort with the two conditions (face-to-face or tape-mediated) as test environments;
- the interaction between the candidate and interlocutor on the live version;
- the adequacy of preparation and response time on the tape version;
- the existence of rater bias with respect to individual candidates on both versions.

Due to the fact that overseas centres cannot always administer both versions of the test, it is likely both versions will survive for the foreseeable future. The most important immediate question, therefore, is not which version is preferable

| 138 access: Issues in language test design and delivery |
| Test-taker performance on versions of the oral interaction module | 139 |
but, rather, how their interchangeability can be enhanced. An important part of
the answer to that question involves minimising the impact of factors other than
the relevant language ability on the test scores of candidates as far as possible.

Of the four most important factors identified above, perhaps the most difficult
to control is the quality of the interaction between candidate and interlocutor
on the live version. This is likely to be even more pronounced in overseas test
centres where the carefully framed guidelines for the selection and training of
interlocutors are not always adhered to as closely as the test developers in
Australia would have hoped. While it should be acknowledged that tighter
recruitment and preparation of interlocutors will not necessarily guarantee the
success of any particular interaction on the live version, it is clearly important
that these procedures be followed as closely as possible. On the basis of this
study, it appears that the interlocutor’s behaviour exerts a powerful influence
not simply over candidates but over raters as well.

The other three factors can be addressed in Australia at the test design and rating
stages. On the basis of the findings in this study, the following changes are
recommended: firstly, longer unassessed warm up periods to allow candidates
to become comfortable in both test environments; secondly, greater preparation
and response times on the tape version; and, finally, the use of a third rater
where either of the two original raters shows evidence of bias towards a par-
ticular candidate.

Of course, there may be other factors related to both the test method and rating
process which account for the lack of consistency in the way candidates are classi-
cified by the two tests. Nevertheless, the insights gained from this investigation
reveal some of the complexity of the factors influencing test scores and how
they interact. In this sense, the findings resonate beyond the particular candidates
studied here.

In addition to limiting the impact of factors other than the relevant language
ability on the test scores of candidates, further investigation into standard setting
needs to be undertaken in conjunction with predictive validity studies to verify
the cut-off scores for each of the key proficiency levels in the test. As suggested
above, it is possible that such research will also improve the degree of equiva-
ence between results on the two versions of the test.

Lastly, the observational, interview and questionnaire data suggest that the
direct and semi-direct versions of the access: oral interaction module may be
tapping two distinct (although overlapping) language abilities — that is, inter-
active versus monologic speaking skills — despite the attempt to ensure parity
between the two versions at the design stage. Of particular importance in this
respect is the role and importance of verbal and non-verbal interaction between
the candidate and interlocutor in the live version which is absent from the tape
version. Even though the candidate’s interactive competence in the live version
is not addressed directly in the scoring criteria, it may still be assessed indirectly
across some or all of these categories. If the two formats do not measure iden-
tical constructs, this is also likely, therefore, to have contributed to the lack of
 correspondence between the test scores found in this study.

In general terms, the results of the study suggest that test scores should be inter-
preted cautiously, particularly when important decisions about individuals are
to be based on them. While language testing continues to strive for a kind of
objective purity or psychometric ‘pristineness’, even in the communicative era,
it appears that test scores on performance tests are produced from a series of
complex interactions between candidates, features of the test format, and raters
in particular. Such a measurement process can hardly claim to be foolproof in
terms of accuracy. In addition, the issue of what exactly is being measured in
performance tests is not always easily resolved but is nevertheless crucial to
the interpretation of test scores. As Spolsky (1995: 358) suggests:

*If the aim [of a language test] is to make some serious decision affecting
the future of the persons being tested, language testers must accept full
responsibility for the inevitable uncertainty of a powerful but flawed
technology, and make sure not just of reliability but also of focused and
relevant validity ...*

Both the direct and semi-direct versions of the access: oral interaction module,
therefore, continue to require rigorous scrutiny in relation both to their reli-
ability and to their validity. This study hopefully demonstrates the importance
and complexity of that task.

**Note**

1. This chapter is a revised version of a paper delivered at the 20th annual
Applied Linguistics Association of Australia (ALAA) congress, Australian
National University, Canberra, September 1995.

Appendix 1: Scoring criteria and descriptors for access: oral interaction module, June 1994

Fluency
7 Speech is marked by a very high degree of fluency.
6 Speech is marked by a high degree of fluency with occasional hesitation.
5 Speech is fluent but with some hesitation or deliberation.
4 Noticeable hesitation and some groping for words is present, but does not impede communication.
3 A marked degree of hesitation, grasping for words or inability to phrase utterances easily impedes communication.
2 Speech is fragmented because of hesitations, pauses or false starts.
1 Fluency is evident only in the most formulaic phrases.

Grammar
7 Range and control of grammatical structures are precise and sophisticated.
6 Candidate uses a broad range of structures with only occasional minor errors.
5 Communication is generally grammatically accurate with a range of structures; minor errors may be noticeable.
4 Satisfactory communication is achieved despite a limited range of structures and/or obvious grammatical inaccuracies.
3 Communication is less than satisfactory because of a limited range of structures and/or the presence of frequent errors.
2 Limited communication is possible but errors are likely to be frequent and intrusive.
1 Severe limitations of grammar prevent all but the most basic communication.
Vocabulary
7 Candidate uses a wide range of vocabulary precisely, appropriately and effectively.
6 Candidate uses a wide range of vocabulary effectively though occasionally may be imprecise.
5 Vocabulary is broad enough to allow the candidate to express ideas well. Circumlocution is smooth and effective, if required.
4 Vocabulary is broad enough to allow the candidate to express most ideas. Can usually circumlocute to cover gaps in vocabulary, if required.
3 Vocabulary is broad enough to allow the candidate to express simple ideas. Circumlocution is sometimes ineffective.
2 Limited vocabulary restricts expression to common words and phrases. Circumlocution is laborious and often ineffective.
1 Vocabulary is very limited.

Intelligibility
7 Speech is clear and can be followed effortlessly.
6 Speech is generally clear and can be followed with little effort.
5 Speech can be followed though at times requires some concentration by the listener.
4 Speech can generally be followed though sometimes causes strain.
3 Speech can generally be followed though frequently causes strain.
2 Speech requires constant concentration to be understood.
1 Speech can only be followed intermittently and then only with considerable effort.

Cohesion
7 Cohesive devices are smoothly and effectively managed.
6 A good range of cohesive devices is used but occasionally these may be inappropriate.
The role of questionnaire feedback in the validation of the oral interaction module

Kathryn Hill

Introduction

Language tests may impact in important ways on people’s lives. They may, for example, influence their ability to enter their chosen course, profession or, as in the case of access, their preferred country of residence. Given such high stakes, it can be argued that the validity of a test needs to be established not only to the satisfaction of test designers but also to that of the group most affected — the test takers themselves. For this reason, a number of researchers, including Shohamy (1982), Zeidner and Bensoussan (1988), Zeidner (1990) and Brown (1991) have recommended the collection of test-taker reactions as a supplement to more formal approaches to test analysis and validation.

While test developers have traditionally sought expert feedback as part of the test validation process, in recent years the effectiveness of feedback from additional sources, most notably from test takers themselves, has been demonstrated by a number of researchers including Alderson (1988); Bensoussan and Zeidner (1989); Brown (1991); Cohen (1984, 1988, 1991); Hill (1994, 1995); Nevo and Sfez (1985); Shohamy (1982); Stansfield (1991); Zeidner (1987, 1988a, 1988b, 1990); Zeidner and Bensoussan (1988).

Such feedback has both practical and theoretical applications and can provide an effective gauge of the face validity of the test to the relevant stakeholders. According to Bachman (1990: 228–229):

*The ‘bottom line’ in any language testing situation ... is whether test takers will take the test seriously enough to try their best, and whether test users will accept the test and find it useful.*
Test-taker feedback can supplement and support statistical analyses by identifying items and aspects of the testing situation which are problematic from the examinees’ point of view. In addition, feedback may indicate ways in which these problems might be resolved.

Feedback from other, more expert additional sources, such as raters and interviewers, also contributes to the evaluation of the validity of a test. This type of feedback may assist in identifying threats to test validity and fairness by shedding light on whether candidates’ language behaviours and test-taking processes are in line with the test designers’ intentions.

The study described in this chapter uses feedback from test takers, raters and interviewers to investigate aspects of validity of the access: oral interaction module from the test takers’ point of view. It examines feedback collected subsequent to the first trial of the access: oral interaction module. While the focus of the study is on test-taker feedback, feedback collected from raters and interviewers is used to support the findings.

The primary focus of this study concerns the contribution of feedback to the validity of the access: oral interaction module in general and the comparability of the live and tape formats in particular. In addition, this study examines whether reactions and performance differ across different types of test taker (male/female; Asian/non Asian). Data for this study was collected during a trial of the oral interaction modules in December 1992 in which test takers took both the live and tape-based formats of the oral interaction module.

Current research
Reactions to oral and written test formats

A number of researchers have focused on how test takers react to different test formats. Zeidner (1990) found university students preferred written to oral examinations. Similarly, Zeidner and Bensoussan (1988) found that, although oral tests were perceived to be more interesting, written tests were perceived to be more ‘pleasant’, ‘valuable’, ‘fair’, ‘reflective of ability’, and to cause less anxiety and be more reflective of the subjects’ comprehension of the stimulus texts. By contrast, Shohamy (1982) found students of Hebrew as a foreign language reacted more positively to oral tests than to cloze tests, regardless of how they performed on the oral test. Taking a different approach, Scott (1986) compared test-taker reactions to an oral EFL proficiency test administered under two different conditions — as a group interview and as a paired interview — and found no significant differences in reactions to the two formats.

How test takers compare live and tape-based test formats is particularly pertinent to this discussion as access: can be administered in either format. Although Stansfield et al. (1990) found the Semi-direct Oral Proficiency Interview (SOPI) — a tape-based test — to be a valid and reliable alternative to the Oral Proficiency Interview (OPI), test-taker feedback revealed that 86 per cent of test takers preferred the live test, 70 per cent felt more nervous on the SOPI, and 90 per cent considered the SOPI more difficult. Respondents were positive, however, about the content and technical quality of the tape-based test as well as about its ability to probe their speaking ability. In a later study which addressed the same issue, Stansfield (1991: 204) found that:

Whilst most examinees prefer the [live format] ... about a quarter of the examinees either have no preference or actually prefer to speak into a tape recorder.

Similarly, Clark (1988) found that although examinees viewed the tape-based tests as generally well constructed, they still felt it was more difficult than the live interview. It has been suggested that the preference for live format tests results from the fact that tape-based formats are less ‘natural’ and therefore more ‘stressful’ (McNamara 1990). However, in McNamara’s study, it may have been the perceived artificiality of the task, rather than test format, that prompted the negative reactions.

Effect of test-taker characteristics on reactions to test

A number of researchers have compared the reactions of different groups of test takers in terms of characteristics such as gender or ethnic background. Zeidner and Bensoussan (1988) found no significant differences between sexes or cultures in attitudes towards written versus oral tests of EFL. However, Bensoussan and Zeidner (1989) found females reacted more negatively and experienced more anxiety than males towards oral language tests. On the other hand, Brown (1991) found no significant differences between the reactions of males and females to an oral and aural test of Japanese language proficiency and Bradshaw (1990) found no gender differences in reactions to a three-part placement test.

In an examination of the effect of ethnicity on reactions, Zeidner (1988b: 73) found no significant difference in attitudes to test format or in mean attitudes...
between ‘western’ and ‘oriental’ respondents, concluding that: ‘sociocultural background variables evidenced as generally weak predictors of test attitudes’.

Effect of test-taker characteristics on test performance

As Bachman remarks, differences in the performance of individuals with different backgrounds and personalities on different types of language test are acceptable only when they reflect genuine differences in the ability being measured. Differences in performance are not acceptable where they appear to be ‘associated with characteristics not logically related to the ability in question’ (Bachman 1990: 221).

A number of studies have looked at the relationship between such test-taker characteristics as gender, language background and performance. Swinton and Powers (1980), for example, found differences in performance between ‘European’ and ‘non-European’ groups on TOEFL. Similarly Alderman and Holland (1981) found significant differences on individual TOEFL items between candidates with the same overall language ability but different native languages. Bensoussan and Zeidner (1989) also found, inter alia, significant differences in performance on an English aptitude test between ethnic groups (Arabic-speaking and Hebrew-speaking) and sexes at an Israeli university. Farhady (1982) found significant relationships between performance on several measures of language ability and sex, university status, academic major and nationality. Spurling and Ilyin (1985) also found that age, first language and educational level significantly affected performance on cloze, reading and listening tests. In contrast, Politzer and McGroarty (1985) concluded that differences in performance across four tests were not uniform across language groups.

Methodology

Test content

The version of the access: oral interaction module which was trialled in December 1992 was designed to assess oral language skills within the range of ‘vocational’ competence only (defined as roughly equivalent to FSI 3, ASLPR 3 or ‘advanced level’). While tasks were designed to have a workplace orientation, task content was not specific to any particular occupation.

As with subsequent versions of the oral interaction test, the specification of task types, the content and the criteria for assessment were designed to be as similar as possible for both the live and tape-based formats. Table 1 compares the content for the two formats.

| Table 1: Specification of content for live and tape-based formats (Dec. 1992 version) |
|-------------------------------------|-------------------------------------|
| Live-interview                     | Tape-based                          |
| Description                        | Description                        |
| Narration (picture sequence)       | Narration                           |
| Exposition 1 (data based)          | Exposition                          |
| N/A                                | Summary (of taped conversation)    |
| Role play 1                        | N/A                                 |
| Role play 2                        | N/A                                 |
| N/A                                | Telephone answering message 1      |
| N/A                                | Telephone answering message 2      |
| Exposition 2 (data based)          | Exposition 2                        |
| Extended discussion                | Extended discussion                 |

Administration of the trial test

The trial was conducted over two days in two separate locations. All candidates attempted both formats of the test. On each occasion half the candidates sat the tape-based format first and half sat the live format first in order to control for any ‘practice effect’.

Subjects

Ninety-four trial candidates, representing a wide range of professional and linguistic backgrounds and age groups, participated in the trials. Candidates were drawn from participants in Adult Migrant English Programs, English language students in Australia on study visas and a group from a work-skills bridging course for overseas qualified permanent residents. All candidates were recent arrivals in Australia and participated on a voluntary basis.

The twelve interlocutors, all of whom were language teachers, interviewed the candidates in pairs. The interviews were taped and the tests were rated at a later time by thirteen raters. These raters were also qualified and experienced language teachers and had received specialist training in rating the test. Questionnaires were administered to each rater at the end of the rating period, and each pair of interlocutors was asked to complete a feedback sheet at the end of the second day of interviewing.
more difficult. The live-interview format, on the other hand, was preferred by the overwhelming majority (83 per cent) who also felt it was a better measure of their ability (90 per cent). This pattern of preferences is remarkably similar to the findings of Stansfield et al. (1990) in relation to the OPI and the SOPI. Table 2 outlines the candidates’ responses.

<table>
<thead>
<tr>
<th></th>
<th>Live (direct)</th>
<th>Tape (semi-direct)</th>
</tr>
</thead>
<tbody>
<tr>
<td>more nervous</td>
<td>40</td>
<td>60</td>
</tr>
<tr>
<td>more difficult</td>
<td>15</td>
<td>85</td>
</tr>
<tr>
<td>preferred</td>
<td>83</td>
<td>17</td>
</tr>
<tr>
<td>better measure of ability</td>
<td>90</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 2: Test-taker reactions to format (N=94)

Parts 1 and 2 of the questionnaire elicited candidates’ reactions to specific aspects of the two tests. Although responses were still consistently more favourable towards the live format, responses to the tape-based format were still positive overall with only three slightly negative reactions (difficulty, preparation and response time). The strength of candidates’ reactions to aspects of the test is summarised in Table 3 using a five-point scale with a score of 1 being the

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>difficulty</td>
<td>live</td>
<td>3.2</td>
</tr>
<tr>
<td></td>
<td>tape</td>
<td>2.9</td>
</tr>
<tr>
<td>fairness</td>
<td>live</td>
<td>3.6</td>
</tr>
<tr>
<td></td>
<td>tape</td>
<td>3.4</td>
</tr>
<tr>
<td>preparation time</td>
<td>live</td>
<td>3.1</td>
</tr>
<tr>
<td></td>
<td>tape</td>
<td>2.9</td>
</tr>
<tr>
<td>response time</td>
<td>live</td>
<td>3.1</td>
</tr>
<tr>
<td></td>
<td>tape</td>
<td>2.9</td>
</tr>
<tr>
<td>clarity of instructions</td>
<td>live</td>
<td>4.3</td>
</tr>
<tr>
<td></td>
<td>tape</td>
<td>3.4</td>
</tr>
<tr>
<td>interviewer helpfulness</td>
<td>live</td>
<td>3.9</td>
</tr>
<tr>
<td></td>
<td>tape</td>
<td>4.5</td>
</tr>
<tr>
<td>accurate measure</td>
<td>live</td>
<td>3.5</td>
</tr>
<tr>
<td></td>
<td>tape</td>
<td>3.0</td>
</tr>
</tbody>
</table>

Table 3: Strength of test-taker reaction

Results

Comparability

Part 3 of the test-taker questionnaire asked:

1 Which test did you feel more nervous in?
2 Which test did you find the most difficult?
3 Which type of test did you prefer?
4 Which test do you think was best able to assess your level of English?

When asked to choose between the two formats, 60 per cent of respondents recorded feeling more nervous in the tape-based test while 85 per cent felt it was
most negative and 5 the most positive reaction. On this scale, a mean score greater than 3 is tending towards positive and less than 3 is tending towards negative. The figures, as shown in Table 3, suggest that the magnitude of the preference for the live format might not be as dramatic as was suggested by the responses to Part 3 of the questionnaire (Table 2).

Correlations between reactions on the respective formats (Tables 4 and 5) help to explain these results. As there was no qualitative data on test-taker reactions, feedback from raters and interlocutors were also useful here. The largest difference in test-taker reactions between the two formats was for clarity of instructions, in favour of the live format (Table 3). The most probable explanation for this difference is the presence of an interviewer, given that the main category of ‘help’ an interviewer could legitimately offer is with clarifying the task. This interpretation is supported by Table 4 which shows that the strongest relationship between reactions to different aspects of the live format is between interviewer helpfulness and clarity of instructions (rho = 0.412). In other words, the more helpful the interviewer was perceived to be, the clearer the instructions were found to be.

Table 4: Correlations between affective reactions (live) (N=94)

<table>
<thead>
<tr>
<th>Difficulty</th>
<th>Fairness</th>
<th>Preparation time</th>
<th>Response time</th>
<th>Instructions</th>
<th>Interviewer</th>
</tr>
</thead>
<tbody>
<tr>
<td>rho</td>
<td>rho</td>
<td>rho</td>
<td>rho</td>
<td>rho</td>
<td>rho</td>
</tr>
<tr>
<td>Fairness</td>
<td>0.203*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prep. time</td>
<td>0.199</td>
<td>0.368***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resp. time</td>
<td>0.210*</td>
<td>-0.105</td>
<td>0.234*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Instructions</td>
<td>0.187</td>
<td>0.230*</td>
<td>0.223*</td>
<td>0.154</td>
<td></td>
</tr>
<tr>
<td>Interviewer helpfulness</td>
<td>0.180</td>
<td>0.270**</td>
<td>0.089</td>
<td>0.047</td>
<td>0.412***</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.236*</td>
<td>0.296*</td>
<td>0.033</td>
<td>-0.208*</td>
<td>0.185</td>
</tr>
</tbody>
</table>

*p<0.05; **p<0.01; ***p<0.001, 2-tailed significance

As well as being less positive than for the live-interview format, there was greater variability in reactions to clarity of instructions for the tape format (Table 3). Despite the steps taken to compensate for the absence of an interviewer on the tape format, raters reported that low level candidates seemed to need more support in understanding the tasks on the tape format, possibly due to the large amount of reading required. It appears from these results that the presence of an interviewer in the live format made a real difference to the perceived clarity of instructions. This may have been because the interviewer could provide some sort of assistance for the weaker candidates.

Table 5: Correlations between affective reactions (tape-based) (N=94)

<table>
<thead>
<tr>
<th>Difficulty</th>
<th>Fairness</th>
<th>Preparation time</th>
<th>Response time</th>
<th>Voices</th>
<th>Instructions</th>
</tr>
</thead>
<tbody>
<tr>
<td>rho</td>
<td>rho</td>
<td>rho</td>
<td>rho</td>
<td>rho</td>
<td>rho</td>
</tr>
<tr>
<td>Fairness</td>
<td>0.312**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prep. time</td>
<td>0.454***</td>
<td>0.345**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resp. time</td>
<td>0.138</td>
<td>-0.094</td>
<td>-0.004</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Voices</td>
<td>0.103</td>
<td>0.362**</td>
<td>0.177</td>
<td>0.145</td>
<td></td>
</tr>
<tr>
<td>Instructions</td>
<td>-0.003</td>
<td>0.007</td>
<td>0.102</td>
<td>0.187</td>
<td>0.121</td>
</tr>
</tbody>
</table>
| Accuracy   | 0.147    | 0.278*           | 0.095         | -0.112 | 0.088        | 0.121

*p<0.05; **p<0.01; ***p<0.001, 2-tailed significance

The majority of respondents perceived the time allowances on the live format to be ‘just right’ (68 per cent for preparation time and 82 percent for response time, compared with 48 per cent and 50 per cent respectively on the tape format). This is not surprising given that on the live format the interviewer could tailor preparation and response time to the individual needs of each candidate. By contrast, the fixed time allowances on the tape-based format may have proved too long for some candidates, and too short for others. A number of candidates told interviewers that they felt hurried on the tape test. Several raters suggested that frustration with time allowances appeared to cause anxiety for some candidates. Raters also reported candidates being cut off mid-sentence on one task, but having too much time on another. Not surprisingly then, there is greater variability in reactions to response time on the tape-based format than on the live format, with marginally more negative reactions to preparation and response time.

This situation may help to explain why high correlations between reactions to preparation time and difficulty were only found on the tape-based format (Table 5). By contrast, the absence of a high correlation on the live format between difficulty and any other variable indicates that there is no single factor which accounts for perceptions of difficulty on this format.
The effects of test-taker characteristics on reactions and performance

Classification of candidates

Candidates were originally classified according to gender, residential status (permanent/temporary), employment status (student/professional), age group (under/over 25) and language background (Asian/European). However, for the purposes of this study, only three categories were analysed: gender, occupation and language background. A cross comparison of membership of these categories is presented in Table 6.

Table 6: Gender and language background by employment status

<table>
<thead>
<tr>
<th></th>
<th>Students</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Female</td>
<td>Male</td>
<td>All</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian L1</td>
<td>9</td>
<td>13</td>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>European L1</td>
<td>1</td>
<td>5</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>10</td>
<td>18</td>
<td>28</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Professionals</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Female</td>
<td>Male</td>
<td>All</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian L1</td>
<td>7</td>
<td>7</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>European L1</td>
<td>14</td>
<td>26</td>
<td>40</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>21</td>
<td>33</td>
<td>64</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In a trial situation, candidates representative of the target population are ideally sought. In this case, however, the actual composition of the group suggests that this may not have been the case for a subsection of the trial cohort. Although the test was designed to assess the general occupational communication skills of prospective migrants, a proportion (34 per cent) of the trial candidates were students and, therefore, ‘non-target-like’ for the purposes of the test.

This raises the issue of whether this group is also unrepresentative of the target population in terms of age and experience, and therefore, whether this group should have been included. However, the inclusion of this group allows us to define language abilities on which the groups could expect to differ. That is, one would predict that professionals would achieve higher scores on a test of vocational English (however vaguely defined) than students. If the two groups perform significantly differently in the manner predicted this could be taken as evidence that the test content and language level is appropriate for the target audience. If, as in Brown’s (1991) study, test takers with occupational experience can be regarded as ‘expert’ in the relevant test domain (in Brown’s case, tour guiding in Japanese) their opinions may also provide evidence for content validity.

Characteristics and reactions

The reactions of males and females, students and professionals and Asian and European background speakers to the following questions were compared:

1. Which test did you feel more nervous in?
2. Which test did you find the most difficult?
3. Which type of test did you prefer?
4. Which test do you think was best able to assess your level of English?

Table 7 indicates significant relationships between language background and degree of nervousness with respect to format. Specifically, Asian language background speakers reported being more nervous on the live format. This finding contrasts with Zeidner (1988b: 73), who found no significant difference in attitudes toward test format between ‘western’ and ‘oriental’ respondents. We may postulate from our findings that candidates from Asian countries may be less familiar with communicative language learning techniques and are therefore more comfortable with the more predictable and structured format of the language laboratory. It is also worth noting that almost half of this group were students who may, therefore, have been relatively familiar with language laboratories.

Table 7: Characteristics and format preference

<table>
<thead>
<tr>
<th></th>
<th>more nervous</th>
<th>more difficult</th>
<th>preference</th>
<th>better measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>gender</td>
<td>1.293</td>
<td>5.212*</td>
<td>1.414</td>
<td>1.129</td>
</tr>
<tr>
<td>occupation</td>
<td>2.703</td>
<td>0.088</td>
<td>0.023</td>
<td>0.033</td>
</tr>
<tr>
<td>L1</td>
<td>6.569*</td>
<td>0.613</td>
<td>0.533</td>
<td>1.783</td>
</tr>
</tbody>
</table>

*p < 0.05

There was also a significant relationship between the format considered most difficult and gender, with female candidates nominating the live interview as...
the more difficult format more often than expected. Given that 85 per cent of candidates felt the tape format to be the more difficult (Table 3), this result is against the trend.6

Table 8 illustrates test-taker characteristics in terms of reactions to the different test formats. These figures were achieved by summing the reactions of each candidate for each format to give an overall ‘affective score’. When Generalised Linear Modelling was used to look at the effect of gender, employment status and language background on reactions, a main effect was found for employment status with ‘professionals’ giving more positive responses than ‘students’ for both formats (live: F=6.12 (1,84), p=0.015; tape: F=11.81 (1,81) p=0.001). In other words, the test was most favourably perceived by the group whose opinion best represents the target population for the test.

Table 8: Summary of reactions (‘all’ vs ‘professionals’)  

<table>
<thead>
<tr>
<th></th>
<th>All (n=94) mean</th>
<th>Professionals (n=64) mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>difficulty live</td>
<td>3.2</td>
<td>3.3</td>
</tr>
<tr>
<td></td>
<td>tape</td>
<td>2.9</td>
</tr>
<tr>
<td>fairness live</td>
<td>3.6</td>
<td>3.8</td>
</tr>
<tr>
<td></td>
<td>tape</td>
<td>3.4</td>
</tr>
<tr>
<td>preparation time live</td>
<td>3.1</td>
<td>3.2</td>
</tr>
<tr>
<td></td>
<td>tape</td>
<td>2.9</td>
</tr>
<tr>
<td>response time live</td>
<td>3.1</td>
<td>3.0</td>
</tr>
<tr>
<td></td>
<td>tape</td>
<td>2.9</td>
</tr>
<tr>
<td>clarity of instructions live</td>
<td>4.3</td>
<td>4.6</td>
</tr>
<tr>
<td></td>
<td>tape</td>
<td>3.4</td>
</tr>
<tr>
<td>interviewer help live</td>
<td>3.9</td>
<td>4.0</td>
</tr>
<tr>
<td></td>
<td>tape</td>
<td>4.5</td>
</tr>
<tr>
<td>voice clarity live</td>
<td>3.5</td>
<td>3.5</td>
</tr>
<tr>
<td></td>
<td>tape</td>
<td>3.0</td>
</tr>
</tbody>
</table>

A comparison of reactions to individual items on the questionnaire shows that the mean responses of ‘professionals’ are not only positive for both formats, but that they are the same as, or more positive than, those of the group as a whole for all reactions except response time (live interview format).

**Characteristics and performance**

The final question addressed in this section is whether performance on the test varies across sub-groups, and what the implications are for test validity of any differences found. Construct validity refers to the ‘adequacy of a test as a measure of the characteristic it is interpreted to assess.’ (Messick 1980: 1012). One approach to construct validation is to compare the performance of groups which might be expected to differ on the construct being measured, for example comparisons between experts and novices (Messick 1988: 55). Whilst differences in performance for gender or language background are ‘construct irrelevant’ for the purposes of this test, differences based on employment status could be considered evidence for construct validity.

Score data for each format was analysed using the multi-faceted Rasch analysis program, FACETS (Linacre 1989–1995) which takes into account candidate ability, item difficulty and rater harshness. This analysis resulted in an ability estimate (expressed as a logit score) for each candidate.7

GLM8 analysis revealed a main effect on test scores for language background (tape: F=5.76 (1,74), p=0.019; live: F=6.93 (1,83), p=0.01) and an interaction effect for language background and employment status (tape: F=8.95 (1,74), p=0.004; live: F=4.81 (1,83), p=0.031). Post hoc tests showed the location of the difference to be between students of Asian and European language background, with the European students outperforming all other groups. The direction of the interaction can be seen from the boxplot (Figure 1), which gives the
mean (+), range (-) and interquartile range (I) of logit scores for each sub-group (tape-based format). The difference between Asian and European professionals, however, was not significant; that is, within the ‘target-like’ group, there was no effect for language background on performance.

Discussion

This study set out to examine the validity of a version of the access: oral interaction module from the test takers’ point of view. Generally, test takers reacted favourably towards both formats of the test. Whilst responses consistently favoured the live format, reactions to specific aspects of the tape-based format still tended to be positive overall. When asked to choose between the two formats, the test takers overwhelmingly favoured the live format. However, these reactions were not uniform across different types of test taker. It is important to note that candidates would not normally have an opportunity to compare the two formats. That is, the possibility remains that reactions to the tape-based format may have been more favourable had respondents not also been exposed to the live format.

The feedback provided some insight into aspects of the test which were problematic for test takers, and suggested directions for improving comparability. One factor contributing to test takers’ perception of the difficulty of the tape-based format was time allowances. This suggested that the time allowances needed to be adjusted very carefully to allow adequate time for weaker candidates to prepare their answers and for stronger candidates to respond in full. The problem does not arise on the live format, where the presence of an interviewer means time allowances can be tailored to individual needs.

Another difference in test takers’ perceptions of the two formats was in clarity of instructions — a difference again attributed to the presence of an interviewer on the live-interview format. Whereas the tape-based format offers the same amount of input for each candidate, raters reported large variations in the amount of help given on the live format. Further constraining the role of the interviewer, therefore, is one means for standardising test-taking conditions between the two formats. However, the role of the interviewer is already severely constrained relative to other live interview test formats (this issue is discussed in depth in Chapter 7).

Further limiting the scope for spontaneity on the live format by controlling interviewer behaviour raises the question of what it is that we are measuring. It has been suggested that research ‘has been on reliability of the ratings perhaps at the expense of the validity of the interview format’ (Ross 1992: 174). Hence, whilst further standardising conditions between the two formats may improve reliability and face validity (in terms of comparability), it should not be allowed to compromise test validity.

While, on the one hand, live interview formats such as the Oral Proficiency Interview (OPI) have been criticised for their ‘inauthenticity’ as spoken interaction (Ross 1992), Pollitt has questioned whether a semi-direct (or tape-based) test format constitutes oral interaction in any sense (comment made to author). According to Stansfield, ‘given the lack of true interaction in the OPI, it is not surprising that the SOPI and the OPI correlate so well’ (1991: 205). O’Loughlin’s (1995) study of the differences in lexical density indicates that the live and tape-based formats of access: are more similar in terms of language output than are OPI and SOPI, suggesting that this is because the live format of access: appears to be less interactive than OPI. Hence, if the interactiveness of the live format is further reduced (eg by further restricting the role of the interviewer), if indeed this is possible, its validity as a test of ‘oral interaction’ is called into question.

The study which has been the subject of this chapter also compared the reactions and performances of different types of test taker. It has been suggested that the opinions of candidates who were non-target-like for the purposes of a vocational English test (ie the students) could be given less weight than those of the more target-like group (ie the professionals) whose opinions best represent those of the target population. Whilst the December 1992 version of access: was not an occupation-specific language test in any practical sense, the workplace orientation of the tasks appears to have been important for face validity. That is, the more favourable responses of the ‘professionals’ (relative to the ‘students’) are an indication that this type of test specificity was important for this group and hence, appropriate. Furthermore, to the extent that this group can be considered ‘expert’ in the domain of vocational English competence, their reactions may also provide evidence of content validity.

Although ‘professional’ candidates reacted more favourably to the test, they did not perform better than the student candidates, apparently weakening the claim of the test to be a measure of vocational competence. However, whilst there was a large difference between the performance of Asian and European students, language background was not a significant factor in the performance of the professionals. That is, this ‘construct irrelevant’ characteristic did not affect the
performance of the target-like group. The fact that these differences disappear within the target-like group could therefore be taken as evidence of test validity.

Conclusion

While greater emphasis is inevitably given to the more conventional methods of comparison, feedback of the kind gathered in this study provides valuable insights into the test takers’ perceptions of the test in general and the comparability of the two formats in particular. While on one level the composition of the trial population was not ideal, the presence of the non-target-like group allowed comparisons between groups which have significance for test validity. It had been anticipated that clear differences would emerge in the performance of the two groups; however, the limited specificity of the test suggests that we need not judge this result too harshly. Indeed, this version of access: could perhaps be more appropriately described as a general proficiency test designed to have face validity for the workplace. That is, a test with this sort of limited specificity could not be expected to differentiate between target-groups and non-target-like groups in the same way as a more occupation-specific test might. On the other hand, the fact that both formats of the test did prove to have high face validity for professionals should not be underestimated. For a high-stakes test such as access: it is important that test-takers should view the test as appropriate for the stated purpose, given that outcomes may have a dramatic impact on the lives of those who take it.

Notes

1. The minimum ‘vocational’ level (Level 5) is described as follows: can communicate effectively in spoken English in a range of social, educational and work situations. Communication is appropriate with a high degree of fluency. Language is grammatically accurate most of the time with a wide range of vocabulary which is used effectively in most situations.

2. Foreign Service Institute scale used to assess the language proficiency of US government employees.


4. European: 18 languages represented; Asian: 15 languages.

5. Cross comparisons of the group revealed that there was a high degree of overlap between ‘professionals’, ‘permanent residents’ and ‘over 25s’, on the one hand, and ‘students’, ‘temporary residents’ and ‘under 25s’ on the other.

6. Whilst there was no indication as to why this might be the case, it is interesting to note that the female subject in O’Loughlin’s (this volume) study (in relation to a later version of the test), whilst preferring the live interview, nevertheless found the tasks on this format slightly harder.

7. Technical problems meant that the actual number of paired (live and tape-based) performances was reduced from 94 to 83.

8. General Linear Model: GLM does analysis of variance and covariance for unbalanced designs. Factors and covariates may be crossed or nested. The model for GLM is a generalisation of the model used in ANCOVA.

9. Whilst the validity of the semi-direct format as spoken interaction has been questioned, it has been suggested that ‘the more extensive language sample [on SOPI] may give it greater content validity [than OPI]’ which represents a single speech style (Stansfield 1991: 203). By virtue of the variety of tasks on each format of access:, the direct format probably avoids the low content validity normally attributed to oral proficiency interviews (eg Shohamy et al. 1986).

References


**Appendix 1: Test-taker questionnaire**

The purpose of this questionnaire is to find out how fair you found the tests.

**Personal information**

1. Please circle FEMALE MALE
2. Please indicate your age range: 16–25 26–35 36–45 46–55 56+
3. How long have you been in Australia? 
4. Do you live here permanently? Yes/No
5. What is your occupation?
6. What is your native language?
7. Do you speak any other languages? Yes/No If yes which ones?

This questionnaire has three parts. In the first part you are asked about the tape test that you took. In the second part you are asked about the interview test that you took. The third part asks about both tests.

Please circle the test which you did FIRST TAPE INTERVIEW

**Part 1: The tape test**

Please tick ONLY ONE answer for each question:

1. Did you feel that the test was:  
   - much too difficult
   - a little too difficult
   - just right
   - a little too easy
   - much too easy
2. Were the questions fair:  
   - very fair
   - quite fair
   - fair
   - a little unfair
   - very unfair

3. Were you given enough time to prepare your answers?  
   - far too much time
   - a little too much time
   - just the right amount of time
   - not quite enough time
   - far too little time
4. Were you given enough time to give your answer?  
   - far too much time
   - not quite enough time
   - just the right amount of time
   - a little too much time
   - far too little time

5. Did you find the voices on the tape clear?  
   - very unclear
   - quite clear
   - clear
   - a bit unclear
   - very clear
6. Were the instructions clear?  
   - very unclear
   - a bit unclear
   - clear
   - quite clear
   - very clear

7. Do you think the test gave a good indication of your English ability?  
   - very good
   - quite good
   - good
   - not very good
   - very poor

**Part 2: The interview test**

Please tick ONLY ONE answer for each question:

1. Did you feel that the test was:  
   - much too difficult
   - a little too difficult
   - just right
   - a little too easy
   - much too easy
2. Were the questions fair:  
   - very fair
   - quite fair
   - fair
   - a little unfair
   - very unfair

---

Access: Issues in language test design and delivery
Appendix 2: Rater questionnaire

Name

Number of tapes assessed
1 Live
2 Tape-based

Please complete this questionnaire AFTER you have completed all rating.

Live version

Section 2

Part A: Describing a building
Agree Disagree
this item elicits a good sample of language
the criteria for the task are appropriate
instructions to the candidate are clear
the level of vocabulary is appropriate
further comments/suggestions (please specify below)

Part B: Cartoon sequence
Agree Disagree
this item elicits a good sample of language
the criteria for the task are appropriate
instructions to the candidate are clear
the pictures are clear and appropriate
further comments/suggestions (please specify below)

Part C: Drowning statistics
Agree Disagree
this item elicits a good sample of language
the criteria for the task are appropriate
instructions to the candidate are clear
the level of vocabulary is appropriate
further comments/suggestions (please specify below)

Part 3: Both tests

Please circle your answer:

1 Which test did you feel more nervous in?
   Tape
   Interview

2 Which test did you find the most difficult?
   Tape
   Interview

3 Which type of test did you prefer?
   Tape
   Interview

4 Which test do you think was best able to assess your level of English?
   Tape
   Interview

Appendix 3: Access: Issues in language test design and delivery

3 Were you given enough time to prepare your answers?
   — far too much time
   — a little too much time
   — just the right amount of time
   — not quite enough time
   — far too little time

4 Were you given enough time to give your answers?
   — far too much time
   — not quite enough time
   — just the right amount of time
   — a little too much time
   — far too little time

5 Were the instructions clear?
   — very clear
   — quite clear
   — clear
   — a bit unclear
   — very unclear

6 Was the interviewer helpful?
   — very clear
   — quite clear
   — clear
   — a bit unclear
   — very unclear

7 Do you think the test gave a good indication of your English ability?
   — very good
   — quite good
   — good
   — not very good
   — very poor
### Section 3

**Role Play 1: Meeting with boss**

<table>
<thead>
<tr>
<th>Agree</th>
<th>Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- this item elicits a good sample of language
- the criteria for the task are appropriate
- instructions to the candidate are clear
- the level of vocabulary is appropriate
- further comments/suggestions (please specify below)

**Role Play 2: Friday lunch**

<table>
<thead>
<tr>
<th>Agree</th>
<th>Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- this item elicits a good sample of language
- the criteria for the task are appropriate
- instructions to the candidate are clear
- the level of vocabulary is appropriate
- further comments/suggestions (please specify below)

### Section 4

**Housing statistics**

<table>
<thead>
<tr>
<th>Agree</th>
<th>Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- this item elicits a good sample of language
- the criteria for the task are appropriate
- instructions to the candidate are clear
- the level of vocabulary is appropriate
- further comments/suggestions (please specify below)

### Section 5

**General questions**

<table>
<thead>
<tr>
<th>Agree</th>
<th>Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- this item elicits a good sample of language
- the criteria for the task are appropriate
- instructions to the candidate are clear
- the questions are appropriate
- further comments/suggestions (please specify below)
Part D: Summary

<table>
<thead>
<tr>
<th>Agree</th>
<th>Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>time is sufficient for the candidate to respond</td>
<td></td>
</tr>
<tr>
<td>this item elicits a good sample of language</td>
<td></td>
</tr>
<tr>
<td>the criteria for the task are appropriate</td>
<td></td>
</tr>
<tr>
<td>instructions to the candidate are clear</td>
<td></td>
</tr>
<tr>
<td>the level of vocabulary is appropriate</td>
<td></td>
</tr>
<tr>
<td>further comments/suggestions (please specify below)</td>
<td></td>
</tr>
</tbody>
</table>

Section 3

Telephone message 1: Travel arrangements

<table>
<thead>
<tr>
<th>Agree</th>
<th>Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>time is sufficient for the candidate to respond</td>
<td></td>
</tr>
<tr>
<td>this item elicits a good sample of language</td>
<td></td>
</tr>
<tr>
<td>the criteria for the task are appropriate</td>
<td></td>
</tr>
<tr>
<td>instructions to the candidate are clear</td>
<td></td>
</tr>
<tr>
<td>the level of vocabulary is appropriate</td>
<td></td>
</tr>
<tr>
<td>further comments/suggestions (please specify below)</td>
<td></td>
</tr>
</tbody>
</table>

Telephone message 2: Party invitation

<table>
<thead>
<tr>
<th>Agree</th>
<th>Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>time is sufficient for the candidate to respond</td>
<td></td>
</tr>
<tr>
<td>this item elicits a good sample of language</td>
<td></td>
</tr>
<tr>
<td>the criteria for the task are appropriate</td>
<td></td>
</tr>
<tr>
<td>instructions to the candidate are clear</td>
<td></td>
</tr>
<tr>
<td>the level of vocabulary is appropriate</td>
<td></td>
</tr>
<tr>
<td>further comments/suggestions (please specify below)</td>
<td></td>
</tr>
</tbody>
</table>

Section 4

<table>
<thead>
<tr>
<th>Agree</th>
<th>Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>University statistics</td>
<td></td>
</tr>
<tr>
<td>time is sufficient for the candidate to respond</td>
<td></td>
</tr>
<tr>
<td>this item elicits a good sample of language</td>
<td></td>
</tr>
<tr>
<td>the criteria for the task are appropriate</td>
<td></td>
</tr>
<tr>
<td>instructions to the candidate are clear</td>
<td></td>
</tr>
<tr>
<td>the level of vocabulary is appropriate</td>
<td></td>
</tr>
<tr>
<td>further comments/suggestions (please specify below)</td>
<td></td>
</tr>
</tbody>
</table>

Section 5

<table>
<thead>
<tr>
<th>Agree</th>
<th>Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>General questions</td>
<td></td>
</tr>
<tr>
<td>time is sufficient for the candidate to respond</td>
<td></td>
</tr>
<tr>
<td>this item elicits a good sample of language</td>
<td></td>
</tr>
<tr>
<td>the criteria for the task are appropriate</td>
<td></td>
</tr>
<tr>
<td>instructions to the candidate are clear</td>
<td></td>
</tr>
<tr>
<td>the questions are appropriate</td>
<td></td>
</tr>
<tr>
<td>further comments/suggestions (please specify below)</td>
<td></td>
</tr>
</tbody>
</table>

Criteria

Please comment on any problems/suggestions you have with the descriptors for each of the criteria.

Fluency

Resources of grammar

Vocabulary

Coherence and cohesion

 Appropriateness

Intelligibility

Comprehension

Overall communicative effectiveness
Interlocutor behaviour (live version)

Please comment on any instances where you felt the performance of the interlocutor negatively influenced the candidate’s performance. Do you have any ideas how this could be avoided in the future?

Any other comments

Appendix 3: Interlocutors’ comment sheet

Name: ____________________________________________________________

Please detail below any comments you have regarding the test.

________________________________________________________________

________________________________________________________________

________________________________________________________________

________________________________________________________________

7

Approaches to the evaluation of interviewer behaviour in oral tests

Janne Morton, Gillian Wigglesworth and Donna Williams

Introduction

A number of highly complex variables interact in oral performance testing and may affect the outcome in terms of the candidates’ scores. A greater understanding of how these variables interact is necessary in order to improve test design and to inform decisions about the training and feedback of those involved at various stages in the cycle. This chapter addresses issues in the reliability and validity of the live version of the access: oral interaction module and investigates two variables related to interviewer behaviour in the oral interaction test. The first is the quality of the interviewer (poor or good) and the second is the nature of the task (structured or unstructured). The study described in this chapter comprised three stages:

1. Good interviewers were differentiated from poor interviewers through rater evaluations of interviewer behaviour.

2. Features of good interviewer behaviour were identified by raters.

3. An analysis of the discourse was undertaken to investigate interviewer performance in structured and unstructured tasks.

Interviewer behaviour

Interviewer performance has been the focus of considerable research. Speech accommodation theory (Beebe and Giles 1984) has shown that while interviewers may facilitate interviews via specific types of questions, modifications or simplifications, they also vary in their ability to facilitate oral interactions (Cafarella 1994 and Filipi 1994). Interviewer language and behaviour in oral performance tests can vary substantially and can influence the language output.
of candidates (Lazaraton and Saville 1994). Listener’s ethnicity may also influence the speaker, who may adjust his or her language accordingly (Ervin-Tripp 1968; Beebe 1983). Young (1995) suggests that an interviewee’s cultural background may also affect his or her perception of roles in an interview and thus the speech output. Young also reports that interviewers’ speech accommodation towards candidates occurs less in highly scripted tests.

Ross and Berwick (1992) examined the characteristics of interviewer accommodation (ie modifications of the form and content of the discourse) in the Oral Proficiency Interview. They found that interviewers systematically varied their discourse to accommodate the learner’s proficiency level, and argued on this basis that at the lower levels of proficiency the extent of interviewer accommodation could serve as a useful measure of proficiency. They recommended that interviewer training should sensitise interviewers to their use of accommodative talk and make them more aware of the cultural context in which the interaction takes place.

Thus, previous investigation of interviewer behaviour indicates two factors relevant to this study. The first is that the interviewer is not ‘neutral’ and the behaviour of the interviewer can, and does, affect the discourse of the candidate. The second concerns systematic cultural differences which have been identified in interviewer behaviour.

The access: oral interaction module: background

The oral interaction module of the access: test is produced in two versions — live interview (which is conducted individually with an interviewer), and tape-based (which is administered to groups of candidates in a language laboratory). Whether the candidate is tested on the tape-based or the live interview is determined by the test centre and depends on several factors: firstly, the number of candidates being tested at each centre, secondly, the technological facilities available, and thirdly, the availability of interviewers. The two versions of the test were developed in parallel, and consist of a wide variety of tasks designed to elicit similar language functions across the two versions. The study which is the subject of this chapter is concerned with the live interview version.

The live version of the oral interaction module includes the following tasks: a description, a narration, an exposition (based on a numerical table or figure), a role-play, and interview questions (see Table 1 in Chapter 6 for a full list of test content). Interviewers are provided with instructions on what to say and how to conduct tasks and are asked to adhere as closely as possible to the script printed in the test booklet. This is in part to ensure that the interview is comparable to the tape-based version, but also because interviewers are selected by the test centre coordinators in the overseas centres, and receive limited training. Tapes of both the live and the tape-based versions are returned to Australia for rating, and each tape is assessed twice by trained raters. In order to provide ongoing monitoring of interviewer behaviour, each rater completes a qualitative questionnaire on the behaviour of the interviewer during the interview, and provides a quantitative assessment of the interviewer’s overall performance on a three-point scale.

Figure 1 (adapted from Kenyon1) shows the interaction of factors which contribute to the score the candidate obtains on the speaking component of the access: test. Note that the interviewer is not the assessor. Thus the rater (assessor) interacts not only with the candidate but also with the interviewer.

Figure 1: Factors in the oral interaction assessment process

Research methodology

Live interview tapes were collected from two administrations that were held in twenty centres in June and July 1994. Three hundred and seventy candidates and sixty-six interviewers were involved. All tapes were rated two or three times by fifty-one trained and experienced raters, as summarised in Table 1. At the time of rating, each rater completed a qualitative questionnaire about the behaviour of the interviewer and then gave the interviewer a global rating on a three point
The test, and Section 5, the final task, are both highly scripted. Section 3 is the least scripted section of the test and involves the interviewer and candidate in a role-play. The transcriptions for these three tasks were then coded to examine interviewer behaviours identified by the raters in their qualitative evaluations. These were:

- the ability of the interviewer to develop rapport with the candidate;
- the ability of the interviewer to manage candidate misunderstandings;
- input and performance across structured versus unstructured tasks.

Results and discussion

This section is divided into three parts. In the first part the results of the quantitative analyses are given. These include statistical evaluations of the raters and the results of the bias analysis. In the second part, the findings of the qualitative evaluations made by the raters are summarised. The discourse analytic approaches in the final section are then used to further investigate interviewer behaviour.

Quantitative analysis

Statistical analysis enabled a ranking of interviewers according to rater perceptions of their competence throughout the interview. Details of the statistical procedures used can be found in Chapter 8. Table 2 presents the ranking for the top ten — the good interviewers (G1–G10) — and the bottom ten — the poor interviewers (P1–P10).

A second analysis — in which the interviewer was included as facet and coded as poor, average or good (the questionnaire appears in Appendix 1). Thus, each interviewer was evaluated a minimum of two times for each interview she or he performed.

Table 1: Participants in research sample

<table>
<thead>
<tr>
<th>Table 1: Participants in research sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Candidates</td>
</tr>
<tr>
<td>Centres</td>
</tr>
<tr>
<td>Interviewers</td>
</tr>
<tr>
<td>Raters:</td>
</tr>
<tr>
<td>Melbourne</td>
</tr>
<tr>
<td>Perth</td>
</tr>
</tbody>
</table>

A total of 761 evaluation questionnaires were completed on the 66 interviewers. Scores for candidates are routinely analysed using the statistical program FACETS (Linacre 1992a, 1992b). (See Chapter 2 for a more detailed description of the FACETS program.) For the purposes of this study, the interviewer scores provided by the raters were also analysed, allowing an evaluation of the interviewer ability which takes into account the harshness or leniency of the rater. Using this technique, a rank ordering of interviewers from poor to good was achieved.

Further quantitative investigation was undertaken with a bias analysis. This was run using the same statistical program, and determined whether there was an interaction between the perceived competence of an interviewer and the scores awarded the candidates. This investigation was to identify any systematic bias observable either from raters in general, or between particular raters and interviewers identified as poor, average or good.

The rank ordering obtained from the statistical analysis was used to identify the top ten and the bottom ten interviewers (ie the ten ‘best’ and the ten ‘worst’) for further investigation. Two approaches, using qualitative data, were employed to investigate the performance of the interviewers. Firstly, an evaluation was made of the characteristics that distinguish good from poor interviewers by collating comments provided by the raters. This was followed by a discourse analysis of the output of these twenty interviewers. Subsequently, for each interviewer two of the candidates they had interviewed were selected. Thus there was a total of forty candidates representing a range of ability levels for transcription and analysis.

Three tasks from the test were transcribed: Sections 2A (picture description), 3 (role-play) and 5 (general questions). Section 2A, the first assessed task in the test, and Section 5, the final task, are both highly scripted. Section 3 is the least scripted section of the test and involves the interviewer and candidate in a role-play. The transcriptions for these three tasks were then coded to examine interviewer behaviours identified by the raters in their qualitative evaluations.
indicates that there is no perceptible difference between the good and average raters. The reliability of separation index provides an estimate of the reliability with which the levels differ from each other. A reliability index of 1 indicates perfect separation. Thus the figure of 0.97 suggests we can be certain about the significance of the separation of levels.

**Table 3:** Interviewer scale measurement report

<table>
<thead>
<tr>
<th>Measure logit</th>
<th>Model error</th>
<th>Interviewer</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.33</td>
<td>0.06</td>
<td>Poor</td>
</tr>
<tr>
<td>0.18</td>
<td>0.02</td>
<td>Average</td>
</tr>
<tr>
<td>0.16</td>
<td>0.02</td>
<td>Good</td>
</tr>
</tbody>
</table>

Separation 5.64  Reliability 0.97

The finding that raters tend to compensate for perceived incompetence in interviewers by awarding candidates higher than expected scores has also been found by McNamara and Lumley (1993). This finding suggests that raters may give the candidate the benefit of the doubt where they consider that the competence of the interviewer is compromised.

An additional variable identified in the present study was the presence of non-native speaking interviewers. The guidelines for *access*: require interviewers to be native English speakers. However, it is the overseas test centre coordinators who select the interviewers and the term ‘native speaker’ appears to be differently interpreted across centres. Table 4 shows the percentage of native English speaking and non-native English speaking (NNS) interviewers (as identified by the raters) for each category; poor, average and good. Half of the interviewers perceived to be poor were NNS, whereas almost all of the average, and all of the good interviewers, were native speakers. The issue of non-native speakers as interviewers is discussed further in the discourse analysis section below.

**Table 4:** Percentage of native speaker and non-native speaker interviewers

<table>
<thead>
<tr>
<th></th>
<th>Poor interviewers</th>
<th>Average interviewers</th>
<th>Good interviewers</th>
<th>Total number of interviewers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Native speaker</td>
<td>5 (50%)</td>
<td>41 (89%)</td>
<td>10 (100%)</td>
<td>56</td>
</tr>
<tr>
<td>Non-native</td>
<td>5 (50%)</td>
<td>5 (11%)</td>
<td>0</td>
<td>10</td>
</tr>
</tbody>
</table>

Thus, the interviewers coded as poor and good were those indicated in Table 2; the interviewers coded as average are the forty-six remaining interviewers who have been omitted from the table. This analysis allowed us to determine the relative ‘difficulty’ of the interviewer. The results are shown in Table 3. The column headed ‘Measure logit’ provides a score for each category of interviewer (ie poor, average or good). This figure represents the degree of difficulty presented by each category of interviewer to the candidate. The negative logit measure for poor interviewers indicates that there is a tendency for candidates to score higher than expected with this category of interviewer, or in other words, that there is a tendency for raters to compensate for what they perceive to be a poor interviewer. The marginal difference between the good and average interviews...
Qualitative analysis

Comments provided by the raters on the ten ‘best’ and ten ‘worst’ interviewers were collated to enable identification of the features which distinguished good from poor interviewers. These are summarised below.

The ability of the interviewer to establish rapport with the candidate was one of the most important aspects that distinguished the two groups of interviewers. Good interviewers tended to interact with candidates, using their names and acknowledging their responses with frequent, varied and supportive feedback. A number of raters commented that the type of feedback given by most good interviewers was more like that in ‘natural conversation’. Poor interviewers demonstrated little evidence of rapport and were frequently felt to be abrupt, brusque or at best neutral. These interviewers rarely gave feedback to candidates. Where feedback was given, it functioned as an ending rather than to encourage further language. Where feedback was lacking, raters found interviewers appeared distant or machine-like rather than involved listeners. Some commented that a tape-mediated version of the test would have been a more reliable instrument.

For low proficiency candidates the development of a supportive, encouraging atmosphere was perceived by raters to be crucial. Although advanced candidates were generally able to perform effectively without interviewer support, low proficiency candidates were perceived to be adversely affected. Comments from raters illustrate this point. Examples 1 and 2 refer to advanced candidates, and example 3 refers to a low proficiency candidate:

1. the interviewer was not very friendly, more like the tape version, however, the candidate was confident so probably no difference (Rater 46);
2. no rapport evident but candidate confident and did not need encouragement (Rater 82);
3. I felt that this candidate would have done a lot better with a more encouraging supportive interviewer to draw out the language (Rater 45).

Modifying prompts and asking additional questions were also considered to be important features of good interviewer behaviour. Substantial variation was found in the way interviewers responded to candidate misunderstandings with good interviewers tending to be more successful in repeating, rephrasing and explaining to ensure that the candidate was given the maximum opportunity to provide an appropriate sample of language. Good interviewers were also more adept at asking additional questions to give the candidates the opportunity to expand their responses. Poor interviewers, on the other hand, were less successful in modifying the prompts where a candidate failed to understand the task or a word in the prompt, or where a candidate gave an irrelevant response. These interviewers also tended to fail to indicate to the candidate that the response was inappropriate. In addition they gave little encouragement to candidates in situations where their responses were very brief and inadequate for accurate assessment.

Differences between poor and good interviewers were most apparent in the role-play. This is the least scripted and therefore most demanding task for the interviewers. Good interviewers generally elicited as much language as possible, often using open-ended questions, encouraging greater language output from the candidate. In contrast, most poor interviewers handled the role-play ineffectively. They either provided little interaction by, for example, failing to give adequate prompts, or less commonly, they dominated, speaking too much. In the worst cases, the interviewers were unable to initiate any dialogue.

In summary, raters assessed good interviewers as having an encouraging, engaged, relaxed style which was responsive to the needs of the candidates. In the case of low proficiency candidates this included slowing speech to adapt to their comprehension level, and rephrasing questions or explaining the task using simpler language. In the case of more proficient candidates this included asking additional questions to extend and challenge the candidate. Generally, the interviewer’s style was seen to benefit the candidate by eliciting the best possible sample of language.

Discourse analysis

The following section provides a finer analysis of the differences between the poor and the good interviewers through an analysis of the discourse. Firstly we investigate whether there are differences in the way good and poor interviewers behave within the interview situation in terms of setting the candidate at ease and creating a relaxed atmosphere. We then focus the analysis on the differences in interviewer performance as a function of the nature of the task. As stated earlier, three tasks were transcribed. Of these, two were structured — the picture description/comparison, and the general discussion section in which the candidate responds to a set of questions. The third task was a role-play. By nature this task is much less structured; the interviewers are provided with
guidelines, but in general may incorporate their own ideas into the task in a way which is not possible in the more structured tasks. The discourse from two interviews from each of the ten best and the ten worst interviewers was analysed.

**The establishment of rapport**

The development of rapport by the interviewer with the candidate is important. In any situation in which a high-stakes test is administered, and in which outcomes can affect the candidate’s future, it is important that interviewers establish a relaxed and encouraging atmosphere for the candidate. Tokens of encouragement (such as ‘well done’, ‘that’s good’, ‘great’, ‘that’s interesting’) were identified in the transcripts to determine whether there were any differences between the two groups of interviewers. Politeness markers (such as ‘please’ and ‘thank you’), modal usage (such as ‘could you…’, ‘would you like to…’), and use of the candidate’s first name were also identified. Given the structured nature of much of the interview, behaviour on the part of the interviewer which assures the candidate that the interviewer is listening is important in creating a relaxed atmosphere. To this end, the use of backchannelling devices were also examined across the tasks. The results of this analysis are given in Table 5.

**Table 5: Markers of rapport**

<table>
<thead>
<tr>
<th></th>
<th>Good interviewers</th>
<th>Poor interviewers</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Encouragement</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>30</td>
<td>11</td>
</tr>
<tr>
<td>Section 3</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Section 5</td>
<td>29</td>
<td>10</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>66</td>
<td>23</td>
</tr>
<tr>
<td><strong>Politeness</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>25</td>
<td>3</td>
</tr>
<tr>
<td>Section 3</td>
<td>18</td>
<td>3</td>
</tr>
<tr>
<td>Section 5</td>
<td>46</td>
<td>40</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>89</td>
<td>46</td>
</tr>
<tr>
<td><strong>Backchannelling</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>179</td>
<td>26</td>
</tr>
<tr>
<td>Section 3</td>
<td>66</td>
<td>41</td>
</tr>
<tr>
<td>Section 5</td>
<td>187</td>
<td>32</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>432</td>
<td>99</td>
</tr>
</tbody>
</table>

The figures in Table 5 show that differences between good and poor interviewers were significant for politeness ($\chi^2 = 16.38, p<.001$) and backchannelling ($\chi^2 = 34.45, p<.001$). All three features were used more by the good interviewers than by the poor interviewers in all sections. This effect is generally more notable in the structured sections (2A and 5) than in the unstructured Section 3. In the structured tasks the interviewer is adopting the role of listener, or facilitator, and the use of these kinds of devices is evidence of active and involved listening on the part of the interviewer. A further point worth noting here is that within the poor group of interviewers the effects were more dramatic for each of the features identified for the non-native speaking interviewers (five were native speakers and five were non-native speakers). In other words, while as a group the poor interviewers used these devices much less than the good interviewers, where these devices were used by the poor interviewers, their use was largely confined to the five native speaker interviewers in this group. One explanation may be that non-native interviewers are not as relaxed in the interview situation as native speaker interviewers. While differences at this level are not the focus of this study, the findings do indicate that further investigation of differences in interviewing behaviour of native and non-native speaking interviewers is warranted.

**Task related differences**

Due to the structured nature of Sections 2A and 5, the scope for interaction with the candidates by the interviewers was limited. However, this was not the case for Section 3 which called upon the interviewer to assume a role in the role-play. The findings from the qualitative questionnaires completed by the raters suggested that for the raters, the performance of the interviewer in the role-play was criterial in their judgment of the competence of the interviewers. Table 6 shows that there was much greater variability in the number of clauses used in Section 3 than was the case in either of the two more structured sections, and that this effect occurred across both groups of interviewers.

**Table 6: Clause average and range by task**

<table>
<thead>
<tr>
<th></th>
<th>Good Average</th>
<th>Range</th>
<th>Poor Average</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Section 2A</strong></td>
<td>382</td>
<td>19.1</td>
<td>331</td>
<td>16.5</td>
</tr>
<tr>
<td><strong>Section 3</strong></td>
<td>713</td>
<td>35.6</td>
<td>568</td>
<td>28.4</td>
</tr>
<tr>
<td><strong>Section 5</strong></td>
<td>252</td>
<td>12.6</td>
<td>232</td>
<td>11.6</td>
</tr>
</tbody>
</table>

There were no significant differences between good and poor interviewers in terms of the number of clauses used in any section, although the poor inter-
viewers used fewer clauses in general. However, the greater range of clauses used by both groups of interviewers in Section 3 suggests that this section allows greater scope for interviewer interaction with the candidates.

Deviations from the script were infrequent in Sections 2A and 5, a trend reflected across both groups of interviewers. Where deviations from the script were made, they tended to be either in response to candidate misunderstandings and/or requests for clarifications, or for the purposes of extending the candidate and eliciting a greater language sample by asking additional questions. One of the differences between the two groups was in the way the interviewers dealt with candidate misunderstanding with the poor interviewers being less successful than the good interviewers. Although the incidence of candidate misunderstandings in this sample was small, the two groups adopted different strategies. Good interviewers were more likely to rephrase, while poor interviewers were more likely to simply repeat what they had said previously. This is shown in Table 7.

**Table 7: Interviewer responses to candidate misunderstandings**

<table>
<thead>
<tr>
<th>Section</th>
<th>Good</th>
<th>Poor</th>
</tr>
</thead>
<tbody>
<tr>
<td>2A</td>
<td>Repeat</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Rephrase</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>Repeat</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Rephrase</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>Repeat</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Rephrase</td>
<td>11</td>
</tr>
</tbody>
</table>

Although the overall difference in distribution of use of repeats versus rephrases is not significant, the difference in distribution is significant in the role-play situation despite the small figures ($\chi^2 = 4.30$, p.<.05).

The role-play task requires candidates to adopt a role in which they are required to make a complaint about a recent incident. The interviewer plays the role of a goods or services provider to whom the candidate must complain. In this task the good interviewers were more likely to adjust their speech input according to the candidate’s level. Interviewer G1 is an example of this. Her first candidate was fluent and confident and the interviewer kept to the script and had no call to repeat or rephrase. Although the resultant interaction was short, the candidate did most of the talking. In contrast, her second candidate was far less confident, requiring the interviewer to play a much more proactive role, including clarifying roles at the outset:

G1: Right, you’ve read the information, and now I’m going to be the manager, and you’re going to be the customer.

The candidate was not always able to grasp what the interviewer said, and she was required to repeat and rephrase at regular intervals so that the role-play could continue. The result was that this interviewer’s input varied considerably across the two candidates.

Poor interviewers were less likely to encourage candidates even when they were having difficulty in this task. For example, with his first candidate, interviewer P2 read the introduction, waited 35 seconds and said ‘Yes sir, can I help you?’. Then, without comment, he allowed the candidate to narrate the incident. Thus the ‘role-play’ was, in fact, a monologue on the part of the candidate — the interviewer uttered the first prompt and from that point on ceased to participate in the role-play at all.

Good interviewers adopted a more facilitative role than the poor interviewers in the role-play. They were more proficient at adapting their techniques to enable each candidate to perform well. Poor interviewers, on the other hand, were perceived by raters to be less at ease in the role-play and were less likely to persist where candidates did not immediately engage with the tasks. One way in which this was manifest was in the lack of additional idea-units used by the poor interviewers, which suggests that they were not fully involved with the task themselves. It is to this issue that we now turn.

In the role-play, the interviewer is required to introduce five idea-units in his or her role as the goods or services provider; that is, the interviewer script presents five ideas for incorporation into the role-play. These ideas are:

- ascertaining the correct location of the incident;
- asking the customer to describe the incident and who was involved;
- asking the customer to explain what compensation he or she is asking for;
- offering an alternative form of compensation;
- finally admitting responsibility by agreeing to the customer’s request.
As we have indicated above, the role-play, due to its unstructured nature, allows the interviewer the greatest degree of flexibility. In other studies of interviewers' behaviour it has been shown that interviewer accommodation plays an important role in the evaluation of candidates (Ross 1992). Raters identified good interviewers as responding better to the candidates as individuals, encouraging negotiation with a confident candidate or playing a low-key part with a weaker one. One way in which this was done was by the introduction of extra idea-units, which added significantly more material to the interaction. In Table 8 the number of scripted ideas used are added to the extra idea-units introduced by the interviewers.

Table 8: Number of idea-units in the role-play

<table>
<thead>
<tr>
<th></th>
<th>First candidate</th>
<th>Second candidate</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Good interviewers</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G1</td>
<td>3</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>G2</td>
<td>9</td>
<td>7</td>
<td>16</td>
</tr>
<tr>
<td>G3</td>
<td>8</td>
<td>5</td>
<td>13</td>
</tr>
<tr>
<td>G4</td>
<td>7</td>
<td>12</td>
<td>19</td>
</tr>
<tr>
<td>G5</td>
<td>6</td>
<td>9</td>
<td>15</td>
</tr>
<tr>
<td>G6</td>
<td>9</td>
<td>4</td>
<td>13</td>
</tr>
<tr>
<td>G7</td>
<td>6</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>G8</td>
<td>3</td>
<td>11</td>
<td>14</td>
</tr>
<tr>
<td>G9</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>G10</td>
<td>0</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>51</strong></td>
<td><strong>64</strong></td>
<td><strong>115</strong></td>
</tr>
<tr>
<td><strong>Poor interviewers</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>P2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>P3</td>
<td>5</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>P4</td>
<td>15</td>
<td>17</td>
<td>32</td>
</tr>
<tr>
<td>P5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>P6</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>P7</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>P8</td>
<td>2</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>P9</td>
<td>1</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>P10</td>
<td>23</td>
<td>18</td>
<td>41</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>46</strong></td>
<td><strong>48</strong></td>
<td><strong>94</strong></td>
</tr>
</tbody>
</table>

Overall, the good interviewers introduced more idea-units into their role-plays than did the poor interviewers, although the difference between the two groups was not significant. However, an examination of the individual differences is revealing. Of the poor interviewers, two (P10 and P4) account for approximately 78 per cent of the extra idea-units introduced by the group of poor interviewers. This suggests that as a group they were reluctant to introduce new material into the role-play. This was not the case with the good interviewers who introduced new material into the role-play for all candidates except where they were interviewing particularly weak candidates who were experiencing great difficulty with the task. Thus, where no extra idea-units are introduced by G9 and G10 with their first candidate, this is the result of the extremely limited command of the language evidenced by these candidates. Since the samples of discourse transcribed represented a stratified sample with a range of candidates at different levels of proficiency, we know this cannot provide an explanation for the dearth of new idea units in the transcripts of the poor interviewers.

The differences between the two poor interviewers who did adopt additional idea-units (P4 and P10) and the good interviewers require discussion. Examination of the transcripts of these interviewers showed that while additional idea-units were indeed adopted, they were not adopted in the spirit of the role-play. P10, in particular, harassed the candidate in her attempts to avoid admitting responsibility for an incident — the candidate played her role valiantly, but there was considerable evidence that the interviewer was making the situation particularly difficult.

The analysis of the discourse suggests that in general the good interviewers are active and involved participants in the interview situation. They are committed to consistently pursuing the best possible outcome for the candidate. On the other hand the poor interviewers appeared uninterested and uninvolved — as the raters commented, they were distant and machine-like. These differences, while evident in all sections of the test, were most pronounced in the unstructured section. The role-play is the most interactive section of the test and allows the interviewer the greatest scope for initiative. Clear and substantial differences were noted in the way the two groups of interviewers approached this task and these differences were consistent across the two groups. Thus the good interviewers adjusted their speech input to the level of the candidate, introduced new ideas where and when appropriate and became involved with the interaction. The poor interviewers, on the other hand, were not as adept at adjusting the content of the task to the level of the candidate and generally avoided the
introduction of new ideas. Where new ideas were introduced this was done in a combative rather than a supportive manner. While the focus of this study has been the performance of the interviewers, and the candidate output has not been examined as a variable, the findings here suggest that in unstructured and more conversational situations the candidate may well be more disadvantaged by a poor interviewer than in the more overtly structured sections of the test.

**Conclusion**

The aim of the project was to identify good from poor interviewers through the raters’ evaluations of the interviewers’ performances and to undertake a qualitative investigation of the differences manifest in the interview technique of good versus poor interviewers. A statistical analysis of the interviewer performances as assessed by the raters allowed us to identify a set of good interviewers and a set of poor interviewers. This information was then used as a basis for qualitative analysis of two kinds. Firstly, the rater questionnaires were examined to identify the features that distinguished the good from the poor raters. Secondly, discourse analysis was undertaken. This revealed differences between the groups in terms of their ability, or desire, to create an encouraging and relaxed environment within which the test could be administered, but it also indicated that it is in the role-play that the differences between the two groups really manifest themselves.

The structured nature of Sections 2A and 5, with their detailed instructions and questions which are read out to candidates, encourages greater uniformity of interviewer behaviour with less likelihood of inadequate technique proving detrimental to candidates (Young 1995). This is an important point since other studies have shown that variability in interviewer discourse can detrimentally affect candidate output (for example, Cafarella 1994; Filipi 1994). Evidence was found that variability in interviewer behaviour may influence the rating of candidates in this study (see Table 3). This is not an isolated incident and has been shown to be the case in other test situations (McNamara and Lumley 1993). The issue of how raters compensate candidates for perceived interviewer inadequacy is one which requires further investigation. In the light of current research on the interviewer as ‘gate-keeper’ (Berwick and Ross 1993), the findings here suggest that in situations in which interviewers have limited or inadequate training, or limited access to training, a highly structured test is likely to minimise any detrimental affects of interviewer technique on scores obtained by candidates.

In the context of the access: test this is an important issue. Interviewers are recruited and trained overseas and test developers have little input into their behaviour. Given that the test is administered in a wide range of centres throughout the world, fairness to all candidates must be a primary concern. The fact that the interview is a structured one would appear from these findings to be a valid choice for this test, since it reduces maximally interviewer variability.

**Note**


**References**


Appendix 1: Rater questionnaire on interviewer behaviour

Please complete one of these sheets for EACH candidate you assess in live mode.

Rater name: __________________ Candidate ID: __________________

Test Centre: __________________

Interviewer details: ID: __________________

Sex: Male □ Female □
Native speaker of English: Yes □ No □

1. Did the interviewer give the candidate:

• appropriate thinking (ie preparation) time? Yes □ No □
• appropriate speaking time? Yes □ No □

If you answered no to either point in question 1, please indicate the effect this had on the candidate.

2. Did the interviewer appear to be sufficiently familiar with the contents of the test booklet? Yes □ No □

3. Did the interviewer provide appropriate prompts for each task as presented in the booklet? Always □ Usually □ Sometimes □ Never □

Do you have any comments about your answer to this question?

4. If the candidate failed to understand the task, what did the interviewer do? (eg repeat the question, rephrase the question, explain vocabulary, etc.)
5. If the candidate didn’t understand or know a word, what did the interviewer do? (eg rephrase, supply the word, etc.)

6. In the case of a completely irrelevant response, what did the interviewer do? (eg allow candidate to continue, bring candidate back to the task — how was this done?)

7. Did the interviewer supply further question prompts where necessary to encourage the candidate to speak more? Please provide details.

8. Did the interviewer provide feedback (eg mm, uhuh, laughs, comments, etc) to the candidate? Please give details.

9. Was the interviewer able to build up a supportive relationship with the candidate? Please provide details.

10. During the role play, was the interviewer effective in eliciting a good sample of language from the candidate?  Yes  No

   Comments:

12. Please comment further on positive interviewer behaviour.

13. Please comment further on interviewer behaviour which was detrimental to the candidate.

14. Do you have any further comments you wish to add?

Please rate the interviewer’s overall performance. Tick one.

<table>
<thead>
<tr>
<th>Poor</th>
<th>Average</th>
<th>Good</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Thank you for your help
A generalisability theory study of ratings and test design in the oral interaction and writing modules

Tim McNamara and Brian Lynch

Introduction

Due to the richness of the assessment context, second-language performance tests introduce a number of factors which may influence the chances of success for a candidate on the test. Figure 1, adapted from Kenyon (1995), offers a schematic representation of aspects, or facets, of the process of performance assessment.

Figure 1: Characteristics of performance assessments

The elicitation and rating of a language performance involves a range of interactions, and each site of interaction introduces a new variable into the process through which a score claiming to represent the candidate’s ability is derived. For example, in speaking tests, the choice of interlocutor is likely to influence
test-retest reliability studies, this approach allows us to specify several sources of error — such as occasions, raters, and items — in a single study. These conditions, or sources of measurement error, are referred to as facets. G-theory allows the investigator to decide which facets will be of relevance to the particular assessment context. A Generalisability study (G-study) can then be designed to estimate the relative effects of these facets using test performance data. This estimation is expressed in terms of variance components, obtained from the mean squares in an analysis of variance where the main effects are persons (the object of measurement in G-theory terminology) and the facets. Variance components are also estimated for the interaction of the main effects. The estimated variance components from the G-study are then used for making decisions about how we may want to specify or modify our measurement procedure. This is done through what is known as a D-study (Decision study), where we can explore ‘what if’ scenarios in relation to the test — what if there were fewer tasks or items? Or more? What if we had fewer ratings, or more? The effect of combinations of these variables (eg one task, two ratings; two tasks, one rating; two tasks, three ratings) on reliability can also be explored. These are summarised in terms of either of two resulting indices of dependability (parallel to classical test theory reliability) under the various conditions we specify.

In this study, we used G-theory to explore the effect on dependability of potential changes to the number of ratings in the oral interaction and writing modules of the test, together with the effect of reducing the number of tasks in the modules. Questions of economy and fairness are involved and these will inevitably be in tension, as we will see.

Generalisability theory

Generalisability theory (G-theory) extends the framework of classical test theory in order to take into account the multiple sources of variability that can have an effect on test scores. A comprehensive introductory account of the theory can be found in Shavelson and Webb (1991) while Bachman (1990) offers a briefer introduction in the context of language testing research. In G-theory the classical test concept of error as being undifferentiated and random is replaced with the identification and estimation of multiple sources of error. Thus, instead of being limited to the estimation of one source of error (eg error across occasion) as in test-retest reliability studies, this approach allows us to specify several sources of error — such as occasions, raters, and items — in a single study. These conditions, or sources of measurement error, are referred to as facets. G-theory allows the investigator to decide which facets will be of relevance to the particular assessment context. A Generalisability study (G-study) can then be designed to estimate the relative effects of these facets using test performance data. This estimation is expressed in terms of variance components, obtained from the mean squares in an analysis of variance where the main effects are persons (the object of measurement in G-theory terminology) and the facets. Variance components are also estimated for the interaction of the main effects. The estimated variance components from the G-study are then used for making decisions about how we may want to specify or modify our measurement procedure. This is done through what is known as a D-study (Decision study), where we can explore ‘what if’ scenarios in relation to the test — what if there were fewer tasks or items? Or more? What if we had fewer ratings, or more?

The effect of combinations of these variables (eg one task, two ratings; two tasks, one rating; two tasks, three ratings) on reliability can also be explored. These are summarised in terms of either of two resulting indices of dependability (parallel to classical test theory reliability) under the various conditions we specify.

In this study, we used G-theory to explore the effect on dependability of potential changes to the number of ratings in the oral interaction and writing modules of the test, together with the effect of reducing the number of tasks in the modules. Questions of economy and fairness are involved and these will inevitably be in tension, as we will see.

Study 1: Oral interaction module

Subjects

The data in this study is from a trial conducted in December 1992 of materials designed for the first administration of the test in April, 1993. The data is from eighty-three trial candidates taking the live version of the oral interaction test. The subjects were drawn from two groups — adult immigrants taking advanced level occupationally or academically related courses on arrival in Australia, and overseas students taking advanced level English language courses prior to commencing study at Australian institutes of higher education. The subjects were paid a small fee to take part in the trial.
Test scoring

The performances of eighty-three candidates were each marked by four raters. For each of seven tasks (picture description, narrative, interpreting data in tabular or graphical form, role-plays and discussion) raters were required to make assessments on a six-point scale for a number of dimensions relevant to that task. For example, in one of the two role play tasks, separate assessments were made for fluency, grammatical resources and vocabulary. In all, the combination of seven tasks and multiple aspects for each task led to a total of twenty-three ratings by each rater for each candidate and these twenty-three rating opportunities were treated as test items in the analysis.

Raters

The four raters in this study were part of a larger group of thirteen raters, all of whom were native speakers of English and trained teachers of English as a Second Language with extensive experience of teaching at advanced levels. In preparation for the first test administration, all raters had attended a one-day training session at which independent ratings of sample tapes were made and discussed, and had subsequently completed the rating of another sample set. Multi-faceted Rasch analysis of the data from these raters was then carried out, and raters meeting predetermined standards for consistency were permitted to participate in the marking of the actual administration. The particular four raters in this study were chosen because of their availability to carry out the enormous amount of rating that was required in this and an associated study. There is no reason to believe that they were in any other way different from the other raters.

Data analysis

The G-study design used in this study was a random effects model with two facets: raters and items. Our universe of admissible observations consisted of trained raters who were native speakers of English, and items that defined characteristics of ESL speaking ability. There were four conditions for the rater facet and twenty-three conditions for the item facet. These facets were considered random in that the four raters and twenty-three items were considered interchangeable with any other set of four raters and twenty-three items from the universe of admissible observations. All analyses were done with the GENOVA program, version 2.2 (Crick and Brennan 1984). In planning our D-studies, we were interested in generalising to twelve combinations of raters and items — all four rater conditions and three sets of item conditions: 8, 16, and 23 items. We were more interested in absolute than in relative decisions; that is, we wanted information that would help us to make decisions about the standing of candidates in relation to a particular standard of performance (the pass mark, and the categories for reporting of results) rather than in ranking the test candidates relative to each other. GENOVA accordingly provides two indices of dependability (parallel to reliability in classical test theory) — the $\phi$ coefficient, reflecting the safety of decisions in relation to some standard such as a pass mark, and the G-coefficient, the equivalent of the reliability index in traditional analysis — which allow us to understand how securely we could rank candidates in relation to each other, but not in relation to a pass standard. G-coefficients are typically higher when raters are involved, as raters may differ in terms of harshness but agree on a ranking. When an absolute standard is involved, on the other hand, harshness becomes relevant and must be taken into account, as it affects the level of agreement with which raters would classify candidates as meeting or not meeting the standard. In what follows, we will concentrate on the $\phi$ coefficient, but also report the G-coefficient.

Results

Tables 1, 2 and 3 consider the effect on dependability of varying the number of ratings given on any single performance. Three different cases are envisioned: using a single rating (Table 1); having two ratings per candidate (Table 2); and having three ratings per candidate (Table 3). In each case we are provided with estimates of the effect of the rater factor on test scores compared with the extent to which test scores reflect actual ability. A percentage of the variability associated with each facet (person, rater, item) is reported for each condition. Table 1 reveals what happens if only a single rating (ie a rating by a single examiner over an entire set of twenty-three items) is provided for each candidate. Fully one-third of the variance in test scores is associated with rater factors, not candidate factors. This is obviously not defensible, and the validity of the test could easily be successfully challenged in a court of law if this were the procedure adopted. The lion’s share of this variability (approximately 25 per cent) is associated with choice of rater, with a further 10 per cent (approximately) reflecting patterns of interaction between particular raters and particular candidates. The Raters facet accounts for the second highest amount of total variance. This means that there is a tendency toward inconsistencies between raters in their judgments; that is, certain raters are more lenient or more severe than others across all
The proportion of total variance associated with candidate ability (the object of measurement) now rises to approximately 78 per cent, approaching (but still falling short of) defensible levels. Over 20 per cent of the variance is still associated with rater factors.

A third rating further improves the situation. As shown in Table 3, the desired proportion of variance has now risen to approximately 84 per cent, with 15 per cent of the variance associated with rater factors. It is obviously not economically feasible for third ratings to be given to all candidates, but these findings support the selective policy of a third rating where the first two ratings are discrepant by more than an agreed amount, or where a narrow fail or narrow pass is involved.

Table 3: Variance components for D-Study No.3 (oral interaction module: 3 ratings x 23 items)

<table>
<thead>
<tr>
<th>Effect</th>
<th>Variance component</th>
<th>Standard error</th>
<th>Percentage of total variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persons (p)</td>
<td>.580</td>
<td>.094</td>
<td>83.6</td>
</tr>
<tr>
<td>Raters (R)</td>
<td>.076</td>
<td>.048</td>
<td>10.9</td>
</tr>
<tr>
<td>Items (I)</td>
<td>.002</td>
<td>.001</td>
<td>0.3</td>
</tr>
<tr>
<td>pR</td>
<td>.029</td>
<td>.003</td>
<td>4.2</td>
</tr>
<tr>
<td>pl</td>
<td>.003</td>
<td>.000</td>
<td>0.4</td>
</tr>
<tr>
<td>RI</td>
<td>.001</td>
<td>.000</td>
<td>0.1</td>
</tr>
<tr>
<td>pRI, e</td>
<td>.003</td>
<td>.000</td>
<td>0.5</td>
</tr>
<tr>
<td>Total</td>
<td>.694</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4 summarises the results of a number of D-studies examining the effect of varying test length and the number of ratings per candidate. An analysis shows that the number of items in the test (the number of aspects of performance rated separately over the different tasks in the test) could be safely reduced; this would achieve economies in rating and in test administration. The difference in the $\phi$ coefficient between having sixteen and twenty-three items with two raters, for example, is .771 and .776, respectively. The results further indicate that having two raters, as opposed to a single rater, results in a marked increase in dependability. For example, the difference between having one rater and having two with twenty-three items is .637 and .776, respectively.
Study 2: Writing module

Subjects

The data in this study is from scripts used as part of the training material for raters of the writing module. These scripts are from actual test candidates taking Version B of the test at a live administration in 1993. The data is from forty-nine candidates.

Test scoring

The performances of the forty-nine candidates were each marked by nineteen raters. Items in the writing test consist of separate ratings of several aspects of performance on three writing tasks (four aspects per task: task fulfilment and appropriacy, conventions of presentation, cohesion and organisation, and grammatical control). The tasks are as follows:

1. (8–10 minutes) Establishing and maintaining interpersonal contacts.
   Example task: Replying to an invitation.

2. (15–20 minutes) Giving/requesting information or explanation.
   Example task: writing a memo.

3. (20–30 minutes) Arguing or discussing an issue.
   Example task: letter to the editor.

Raters

The nineteen raters in this study were all native speakers of English and trained teachers of English as a Second Language with extensive experience of teaching at advanced levels. All had attended a one-day training session at which independent ratings of sample scripts were made and discussed, and had subsequently completed the rating of the set of forty-nine scripts. Six weeks later, as part of a study on feedback to raters using the bias analysis feedback in multi-faceted Rasch measurement (Wigglesworth 1993), the raters carried out ratings of the forty-nine scripts again. This data was used in the analysis.

Data analysis

The G-study design used in this study was a random effects model with three facets: raters, tasks and items. Our universe of admissible observations consisted of trained raters who were native speakers of English, tasks that represented general purpose writing tasks for adult citizens of Australia in their private and public lives, and items that defined characteristics of ESL writing ability. There were nineteen conditions for the rater facet, three conditions for the task facet and four conditions for the item facet. These facets were considered random in that the nineteen raters, three tasks and four items were considered interchangeable with any other set of raters, tasks and items from the universe of admissible observations. All analyses were done with the GENOVA program, version 2.2 (Crick and Brennan 1983). D-studies were requested for eighteen combinations of raters, tasks and items — up to three rater conditions, up to three task conditions and two sets of item conditions: 3 and 4 items. As before, absolute decisions (ϕ coefficients) were of greater interest, although Generalisability coefficients are reported as well.

Results

Tables 5 and 6 report the extent of variability with each of the facets of the assessment setting examined in the study — raters, items and persons (the latter being the object of measurement). Each table envisions a different scenario, depending on the number of ratings involved.

### Table 4: Generalisability coefficients for the twelve oral interaction module D-Studies

<table>
<thead>
<tr>
<th>Number of raters</th>
<th>Number of items</th>
<th>G-coefficient</th>
<th>ϕ-coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>23</td>
<td>.955</td>
<td>.870</td>
</tr>
<tr>
<td>3</td>
<td>23</td>
<td>.943</td>
<td>.836</td>
</tr>
<tr>
<td>2</td>
<td>23</td>
<td>.919</td>
<td>.776</td>
</tr>
<tr>
<td>1</td>
<td>23</td>
<td>.853</td>
<td>.637</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
<td>.952</td>
<td>.866</td>
</tr>
<tr>
<td>3</td>
<td>16</td>
<td>.938</td>
<td>.832</td>
</tr>
<tr>
<td>2</td>
<td>16</td>
<td>.913</td>
<td>.771</td>
</tr>
<tr>
<td>1</td>
<td>16</td>
<td>.846</td>
<td>.632</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>.939</td>
<td>.851</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>.925</td>
<td>.816</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>.897</td>
<td>.755</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>.823</td>
<td>.615</td>
</tr>
</tbody>
</table>
In the case of a single rating of each candidate (Table 5), an unacceptably high proportion of the variance is associated with the combined effect of rater differences (12.4 per cent of the variance) and rater by person interactions (8.6 per cent of the variance). A further component of the variance is associated with task-related interactions — almost 4 per cent of the variance is associated with Person x task and Person x rater x task interactions. The situation is not as problematic as for the oral interaction test, but it is not acceptable.

**Table 5: Variance components for D-Study No.1 (writing module: 1 rating x 3 tasks x 4 items)**

<table>
<thead>
<tr>
<th>Effect</th>
<th>Variance component</th>
<th>Standard error</th>
<th>Percentage of total variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persons (p)</td>
<td>1.027</td>
<td>0.213</td>
<td>73.53</td>
</tr>
<tr>
<td>Raters (R)</td>
<td>0.173</td>
<td>0.057</td>
<td>12.38</td>
</tr>
<tr>
<td>Tasks (T)</td>
<td>0.003</td>
<td>0.003</td>
<td>0.22</td>
</tr>
<tr>
<td>Items (I)</td>
<td>0.001</td>
<td>0.001</td>
<td>0.07</td>
</tr>
<tr>
<td>pR</td>
<td>0.120</td>
<td>0.008</td>
<td>8.57</td>
</tr>
<tr>
<td>pT</td>
<td>0.025</td>
<td>0.004</td>
<td>1.76</td>
</tr>
<tr>
<td>pI</td>
<td>0.002</td>
<td>0.000</td>
<td>0.15</td>
</tr>
<tr>
<td>RT</td>
<td>0.001</td>
<td>0.000</td>
<td>0.04</td>
</tr>
<tr>
<td>RI</td>
<td>0.002</td>
<td>0.000</td>
<td>0.14</td>
</tr>
<tr>
<td>TI</td>
<td>0.000</td>
<td>0.000</td>
<td>0.00</td>
</tr>
<tr>
<td>pRT</td>
<td>0.027</td>
<td>0.001</td>
<td>1.91</td>
</tr>
<tr>
<td>pRI</td>
<td>0.005</td>
<td>0.000</td>
<td>0.52</td>
</tr>
<tr>
<td>pTI</td>
<td>0.001</td>
<td>0.000</td>
<td>0.04</td>
</tr>
<tr>
<td>RTI</td>
<td>0.000</td>
<td>0.000</td>
<td>0.01</td>
</tr>
<tr>
<td>pRI, e</td>
<td>0.009</td>
<td>0.000</td>
<td>0.64</td>
</tr>
<tr>
<td>Total</td>
<td>1.396</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The situation improves considerably when there is a second rating of each candidate (Table 6). Now the variance associated with the combined effect of rater differences (7 per cent of the variance) and rater by person interactions (4.9 per cent of the variance) has been reduced, with almost 84 per cent of the total variance accounted for by the ability of the candidates (the object of measurement). The effect of interactions involving task has been reduced to approximately 3 per cent. The figures for rater-related variance may be interpreted in this way: the averaging of scores from the two ratings reduces the amount of variance associated with a single rating; what is left is the ‘luck of the draw’ about which two raters a candidate may get — they could both be relatively harsh, for example, or relatively lenient — in which case there will be a clear effect on the likely outcome for candidates in terms of the criterion performance.

**Table 6: Variance Components for D-Study No. 2 (writing module: 2 ratings x 3 tasks x 4 items)**

<table>
<thead>
<tr>
<th>Effect</th>
<th>Variance component</th>
<th>Standard error</th>
<th>Percentage of total variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persons (p)</td>
<td>1.027</td>
<td>0.213</td>
<td>83.66</td>
</tr>
<tr>
<td>Raters (R)</td>
<td>0.086</td>
<td>0.029</td>
<td>7.90</td>
</tr>
<tr>
<td>Tasks (T)</td>
<td>0.003</td>
<td>0.003</td>
<td>0.25</td>
</tr>
<tr>
<td>Items (I)</td>
<td>0.001</td>
<td>0.001</td>
<td>0.08</td>
</tr>
<tr>
<td>pR</td>
<td>0.060</td>
<td>0.004</td>
<td>4.88</td>
</tr>
<tr>
<td>pT</td>
<td>0.025</td>
<td>0.004</td>
<td>2.00</td>
</tr>
<tr>
<td>pI</td>
<td>0.002</td>
<td>0.000</td>
<td>0.17</td>
</tr>
<tr>
<td>RT</td>
<td>0.000</td>
<td>0.000</td>
<td>0.02</td>
</tr>
<tr>
<td>RI</td>
<td>0.001</td>
<td>0.000</td>
<td>0.08</td>
</tr>
<tr>
<td>TI</td>
<td>0.000</td>
<td>0.000</td>
<td>0.00</td>
</tr>
<tr>
<td>pRT</td>
<td>0.013</td>
<td>0.001</td>
<td>1.09</td>
</tr>
<tr>
<td>pRI</td>
<td>0.004</td>
<td>0.000</td>
<td>0.29</td>
</tr>
<tr>
<td>pTI</td>
<td>0.001</td>
<td>0.000</td>
<td>0.04</td>
</tr>
<tr>
<td>RTI</td>
<td>0.000</td>
<td>0.000</td>
<td>0.01</td>
</tr>
<tr>
<td>pRI, e</td>
<td>0.004</td>
<td>0.000</td>
<td>0.36</td>
</tr>
<tr>
<td>Total</td>
<td>1.227</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 7 reports the coefficients resulting from varying the number of ratings, tasks and items. It can be seen that the reliability under operational conditions — two raters, three tasks, four items; absolute (ie criterion-referenced) decisions — is acceptable: $\phi = .837$. It also shows us what effect alternatives to current procedures would have. These alternatives involve either varying the structure of the test so as to include fewer tasks and/or items or varying the number of raters required to judge each candidate’s work.

**Modification of the test structure**

Modification of test structure involves either or both of the following:

- Reducing the number of tasks, for example from three to two, or even to one. This would save considerably on test time and on rating time.
Reducing the number of aspects rated on each task. This would save some time in the rating process, although it would not save test time, and is thus of less practical value.

Table 7: Generalisability and $\phi$ coefficients for the eighteen writing module D-studies

<table>
<thead>
<tr>
<th>Number of ratings</th>
<th>Number of tasks</th>
<th>Number of items</th>
<th>G-coefficient</th>
<th>$\phi$-coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
<td>.760</td>
<td>.667</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>3</td>
<td>.836</td>
<td>.773</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>3</td>
<td>.865</td>
<td>.817</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>4</td>
<td>.767</td>
<td>.673</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>4</td>
<td>.841</td>
<td>.778</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>4</td>
<td>.869</td>
<td>.820</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>.818</td>
<td>.714</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>3</td>
<td>.884</td>
<td>.818</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>3</td>
<td>.909</td>
<td>.859</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>4</td>
<td>.823</td>
<td>.719</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>4</td>
<td>.889</td>
<td>.821</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>4</td>
<td>.911</td>
<td>.862</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>3</td>
<td>.840</td>
<td>.732</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>3</td>
<td>.902</td>
<td>.834</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>.924</td>
<td>.874</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>4</td>
<td>.844</td>
<td>.735</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>4</td>
<td>.904</td>
<td>.837</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>4</td>
<td>.927</td>
<td>.877</td>
</tr>
</tbody>
</table>

Reducing the number of tasks

Table 8 demonstrates the consequences of reducing the number of tasks. If we hold constant the number of ratings (2) and the number of items per task (4), the dependability coefficients for varying numbers of tasks is as reported in Table 9. This data shows that there is a considerable gain in increasing the number of tasks from one to two, and a smaller gain in a further increase from two to three tasks. This means that there may be an argument (although by no means a conclusive one) for reducing the number of tasks in the test from three to two without seriously jeopardising the dependability of candidates’ scores.

Table 8: Effect on dependability of varying the number of tasks
(assuming two ratings and four items per task)

<table>
<thead>
<tr>
<th>Number of tasks</th>
<th>$\phi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>.837</td>
</tr>
<tr>
<td>2</td>
<td>.821</td>
</tr>
<tr>
<td>1</td>
<td>.773</td>
</tr>
</tbody>
</table>

If a decision were made to remove a task, the question arises as to which would be the best one to remove. Task 1 or Task 2 are sufficiently similar to suggest that one or other of these might be removed — possibly Task 2, as Task 3 also requires writing for the public domain. On the other hand, there may be an argument in terms of content validity (and even construct validity) for the inclusion of a range of writing tasks. This, however, may not be crucial given the potential range of such tasks facing candidates in the real world and the difficulty of specifying which of them will be relevant to candidates. Clearly, some idea of general proficiency in writing underlies the approach of the test.
the key to achieving an acceptable level of dependability is the number of ratings of any candidate.

Table 9: Effect on dependability of varying the number of items (assuming two ratings per item)

<table>
<thead>
<tr>
<th>Number of tasks</th>
<th>Number of items</th>
<th>ϕ</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
<td>.834</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>.818</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>.773</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>.837</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>.821</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>.778</td>
</tr>
</tbody>
</table>

Varying the number of ratings

Table 10 reproduces part of the information from the relevant D-studies reported in Table 7. It shows the effect on dependability of varying the number of ratings for tests of varying structure (assuming that each task has four items, as at present). As can be seen, the use of a single rater produces unacceptably low levels of dependability ($ϕ = .719$ for two tasks, $ϕ = .735$ for three tasks); the use of two raters, as at present, seems amply justified by the data. The resulting dependability estimate is publicly defensible ($ϕ = .837$), perhaps at the lower level of confidence (estimates of .9 or better would be preferable). The practice of three ratings per candidate would come closer to this level (for three raters and two tasks, $ϕ = .862$; for three raters and three tasks, $ϕ = .877$). However, this may not be realistic on economic grounds. These findings endorse the following practices in the rating of the writing module:

- where there is a substantial discrepancy between any two raters, a third rating should be made;
- in borderline cases (close to the pass point on the test), a third rating should be routine.

These two practices effectively secure the gains in dependability that the analysis suggests are appropriate.

It is worth noting that gains in dependability are more readily achieved by the addition of another rating than by the addition of another task (i.e., by maintaining or extending the length of the test), as shown in Table 10. A shorter test of, say, two tasks with double ratings ($ϕ = .821$) is more dependable than a longer test (three tasks) with a single rater ($ϕ = .735$). As with the oral interaction test, the key to achieving an acceptable level of dependability is the number of ratings of any candidate.

Table 10: Effect on dependability of varying the number of ratings (assuming 4 items per task)

<table>
<thead>
<tr>
<th>Number of ratings</th>
<th>Number of tasks</th>
<th>ϕ</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
<td>.877</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>.837</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>.735</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>.862</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>.821</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>.719</td>
</tr>
</tbody>
</table>

Conclusion

This study has confirmed the usefulness of Generalisability Theory in making decisions about test design. Using data from an administration of the test, it enables us to investigate the contribution to the awarded test score of aspects of the test situation which are either desirable (candidate abilities) or undesirable (rater factors, interactions between test candidates and raters, and interactions between raters and tasks, or raters, persons and tasks). The analysis also helps us to envisage the effect of modifications in test length.

The results reveal that the current practice of double rating of performances on the speaking and writing module is absolutely crucial to the defensibility of the ratings made, and that third ratings in borderline or discrepant cases are justified. The use of single ratings cannot be defended on the basis of this data. On the other hand, the analysis supports an argument in favour of some reduction of the length of the oral interaction module, and also possibly of the writing module. This would save test administration time, and also achieve economies in the rating of performances from audiotape and/or written script.

Notes

1. This research was carried out in the National Languages and Literacy Institute of Australia (NLLIA) (now Language Australia) Language Testing Research Centre (LTRC) in the Department of Applied Linguistics and
Language Studies at the University of Melbourne. The original contract for this research was awarded to the first author, but the second author was involved from the outset, and his expertise in Generalisability Theory was crucial to the success of the project. We also wish to acknowledge the help of Kieran O’Loughlin, Gillian Wigglesworth, Tom Lumley and Janne Morton of the NLLIA LTRC for their assistance in several aspects of the project.

2. One complication in our design was the fact that items on the access: test are associated with certain tasks. Because of the balanced design requirements for GENOVA, it was not possible to consider the facet of task, which may have been considered as a fixed effect and within which items would have been nested.

3. Originally 50 scripts were used, but data was incomplete from one script, and it was excluded from the analysis as GENOVA cannot handle missing data.

4. Originally data from 20 raters was involved, but data was incomplete from one rater, and it was excluded from the analysis as GENOVA cannot handle missing data.

References


9

Text type and rater decision-making in the writing module

Susan Delaruelle

Introduction

In assessing second language writing it is common to use analytic scales that describe the key features of the performance elicited in the test. As Pollitt and Murray (1996: 76) point out, during the assessment process raters must be able to easily match the performance to the set of descriptors they use. However, this raises the question of whether it is appropriate to use scales containing generic criteria that are applied to whatever kind of text is produced (thus possibly requiring raters to ignore features of a text that are particularly salient to a given text type) or whether there is a case for having task-specific criteria for different kinds of writing tasks.

This chapter sets out to address these questions by investigating the extent to which raters’ decisions concerning the salience of the rating criteria used in the access: writing module are influenced by the type of text being rated. A secondary aim of the study is to explore differences in the rating behaviour of inexperienced and experienced teachers.

The access: writing module rating procedure

The writing module of the access: test seeks to sample candidate performance on a range of general purpose tasks which relate to various authentic contexts. These contexts reflect those encountered by adults in Australia as they interact in English in both the public and private domains.

During the one-hour test, candidates are required to complete three tasks which relate to three distinct areas of functional purpose. The first task emphasises interpersonal communication and has the social purpose of establishing or
determine the salient features of that text type as a basis for the development of a scale. This process enables raters to assess what is salient in future rating sessions (see, for example, Hamp-Lyons 1991c).

Clearly, in the development of a multiple-trait instrument, the process by which the salient criteria for a given text type are determined is critical if the scale is to be valid. One method of determining which features are most salient to a particular text type is to have raters read and assess sample texts in the absence of predetermined criteria to see which features commonly emerge as being most salient. Thus the raters react to the way candidates have responded linguistically to the particular requirements of the test task and establish which are the most salient features of specific tasks. It is this method which is used in the following study.

**Multiple-trait scoring**

These three tasks are then rated using a holistic method known as multiple-trait scoring. This procedure involves the scoring of a written text on several traits, or facets, exhibited by the text. All three tasks are rated against the same performance criteria — the scoring procedure thus assuming that it is appropriate to use the same assessment criteria and performance descriptors to assess all texts, irrespective of task type. The criteria, or traits, selected as being most appropriate for the assessment of the tasks on this test are task fulfilment and appropriacy, conventions of presentation, cohesion and organisation, and grammatical control. Each of these criteria is marked on a scale from 0–5 using performance descriptors at each level.

One disadvantage of multiple trait scoring is that raters take longer to evaluate three or four individual traits than they would to give a holistic score, and it is therefore a more expensive procedure. However, there is a trade-off in reliability and validity (Hamp-Lyons 1990). With holistic scoring, reliability is achieved by a number of raters reaching consensus on the quality of the text. With multiple-trait scoring, the reliability of the instrument is increased by the greater number of scores available for a single piece of writing. As reliability is generally linked to the number of candidate performances assessed (Weir 1993), the more evidence available about a candidate's ability, the more confidence we may have about our judgments concerning this ability. In this regard Hamp-Lyons (1990: 79) states:

... the rationale generally given for multiple scoring is that multiple judgments lead to a final score that is closer to a 'true' score than any single judgment.

The validity of a multiple-trait scoring procedure will naturally depend on the rigour with which the scoring instrument is developed. First, a careful analysis of the written requirements of the discourse community must take place, then prompts need to be designed to elicit the type of writing required in that particular context. Working within this context, test developers and raters then determine the salient features of that text type as a basis for the development of a scale. This process enables raters to assess what is salient in future rating sessions (see, for example, Hamp-Lyons 1991c).

**Task variability**

The two writing tasks used in this study were different in terms both of their functional purpose and their topic. It was expected that these tasks would place quite different linguistic demands on test candidates and that this would be reflected in their performance. Other researchers have examined the effect these two aspects of a task have on candidates' test scores.

Bachman (1990), Hamp-Lyons (1990), and Weir (1993), for example, have argued that a variety of factors, either alone or in combination, may affect a writer's test score. These include the type of written task the candidate is required to produce, the topic of a writing task, and the amount of information included in the task prompt. The nature of the written text may disadvantage some candidates while advantaging others. Bachman (1990), for example, suggests that unfamiliarity with a particular text type may make the expected response more difficult for test candidates. In addition, the task or genre requirements of a text will dictate to a certain extent the grammatical and lexical resources employed at the sentence level, while at the discourse level they will influence choices concerning textual organisation and layout, and the use of cohesive devices.

The issue of test task variability has been examined in the context of the Test of Written English (TWE), the writing module of the Test of English as a Foreign Language (TOEFL). The test is designed to determine candidates' ability to cope with the rigours of academic study. Bridgeman and Carlson (1983) initially identified two types of writing tasks that best represented what prospective students would be expected to produce in tertiary settings, namely a report task and a
They hypothesised that these two tasks would elicit different writing skills. However, this hypothesis was not supported in a study by Carlson et al. (1985). As part of the piloting of tasks for the TWE, the researchers asked students to complete four writing samples which were then holistically rated. The samples included two discussion tasks and two report tasks. They found no significant difference in the way the different text types ranked students. They suggested that some of this apparent lack of difference between tasks could have been a result of rater behaviour and some aspects of the scoring procedure. They also suggested that the raters may have experienced a halo effect where the overall qualities of the first text they read influenced the ratings of the subsequent texts by the same candidate.

Reid (1990) used the Writers Workbench (WWB) Stylefiles1 program to analyse a sample from the same pilot TWE corpus to evaluate the impact of task on writers’ responses in the assessment process. Reid compared the written products generated by two report tasks and the two discussion tasks to determine any significant linguistic differences between the two task types and the four task topics. The syntactic variables she found included average sentence length, percentage of short sentences, percentage of complex sentences and percentage of passive-voice verbs. Earlier research had suggested that more investigation was needed into the kinds of structures that make up sentences, and that there may well be differences too discrete — or too complex — for the WWB analysis to quantify (Faigley 1980).

In her study, Reid also noted that the reports generated longer texts and hypothesised that this was an indicator of greater writing fluency. The report texts also contained considerably longer words than the discussion texts, indicating more formal, informational responses than the informal, interpersonal responses to the discussion genre. However, these findings may reflect another variable; that is, the actual amount of information given in a task prompt (which can vary greatly — from a single sentence to a paragraph, or to one which includes graphical or diagrammatic information). Although Reid found that the report prompts generated lengthier texts than the discussion prompts, the report prompt also contained a lot of information which, if copied, could have lengthened texts considerably. She suggests this might explain the high incidence of longer words in the report texts. Reid, however, does not report on the relationship between topic differences and test scores.

The topic of a task can significantly affect the content and quality of an essay. Park (1988) used a small subsample of the Carlson et al. (1985) corpus to study the effect of topic on test performance and found that Chinese and English students majoring in hard science subjects received markedly higher scores on one of the report tasks (on the topic of farming) than one of the discussion tasks (on the topic of space). However, the scores of students doing social science majors did not differ significantly across the two writing topics. The observed differences between scores for the former group of students may indicate greater familiarity with the report genre due the area of educational specialisation of these students, rather than obvious differences in topic.

Despite these findings, however, an investigation into the variability associated with test tasks has led Hamp-Lyons (1991a: 103) to assert that prompt difficulty does not reside so much in the prompt, but in the person rating the text. She writes:

'It has led me back to issues of the impact of readers on texts and judgements of text quality, so that … I realise how improbable it is that we will ever pin down prompt difficulty, since so much of the difficulty is not in the prompt or the writer, but in the reader.'

Rater behaviour

Raters vary in their approach to scoring written texts in terms of their harshness, leniency and consistency (Jacobs et al. 1981; McNamara 1996). This variability may reflect differences in raters’ perceptions of what constitutes quality in writing (Jacobs et al. 1981; Hamp-Lyons 1991a). This study focuses on whether the qualities valued by raters in a piece of writing vary according to the type of task or whether raters value certain qualities irrespective of task type. A number of other studies have investigated this issue.

In one of the first studies of its kind (Diederich, French and Carlton 1961) fifty-three raters judged 300 expository essays (150 essays each of two topics) written by L1 college freshmen. The raters were asked to annotate each of the 300 essays they assessed. These annotations were then classified and, together with the ratings, were factor analysed. The analysis revealed five clusters of qualities on which raters based their assessments. These were ideas (relevance to the topic and the writer’s purpose, clarity, quantity, development and persuasive-
Hamp-Lyons (1991b) investigated what four trained experienced ELTS raters valued in written texts as they read, scored and discussed twenty-three ELTS essays. She hypothesised that if raters valued content, as did subject specialists, this would emerge through the discussion of the essays, although there was no criterion related to content in the ELTS scoring guide. Hamp-Lyons looked at those parts of the data which addressed issues of content; that is, quality of content, specialist vocabulary, and relevance. She found very little emphasis on content, nor did the raters value specialist vocabulary. She also found that the raters took a very narrow view of what were relevant criteria. Hamp-Lyons (1991b: 143) concludes:

When raters are rating the ELTS writing test they respond from their own set of values, and do not try to respond to student writing as they think subject specialists would.

Thus, both first language and second language research indicates that there is no clear agreement amongst raters as to which qualities they value most in writing. On the one hand there seems to be a large body of research which suggests that when rating holistically, raters are generally influenced by the content and organisation of texts. However, studies by Deiderich et al. and Harris suggest that some raters are influenced by errors in usage, sentence structure, punctuation and spelling. Other studies (Stewart and Grobe 1979; Charney 1984) suggest that raters using holistic scoring procedures often equate quality in writing with features such as length, script and lack of spelling errors. The Bridgeman and Carlson (1983) and Hamp-Lyons (1991b) studies point to the necessity for test developers to ensure that writing tasks and scoring procedures reflect what is required in the specific context in which the students/candidates will find themselves.

Protocol analysis

One approach to an investigation of the effects of task type on rater behaviour is through the analysis of raters’ verbal reports or protocols. During the collection of think-aloud protocols, raters verbalise their thoughts while rating a number of texts. The verbal reports are recorded, transcribed and then analysed according to a coding system. Carefully elicited and interpreted, verbal reports may be a valuable and reliable source of information about cognitive processes of the informant (Ericsson and Simon 1984). A number of studies have investigated various aspects of rater behaviour.
Huot (1990) analysed the protocols of four trained, experienced raters and four untrained, inexperienced raters who used a holistic scale to assess forty-two student essays. Huot found no differences in the rating criteria for the two groups but he did find that the experienced raters had a broader range of responses to the essays and had more efficient rating strategies overall.

Cumming (1990) analysed the protocols of expert and novice raters assessing student essays which varied on two levels, ESL proficiency (intermediate and advanced) and writing expertise in the mother tongue (average and professional). The essays were rated on a four-point scale for their effectiveness in terms of ‘language use’, ‘rhetorical organisation’ and ‘content’. The study showed that the novices were far more lenient than the experienced group on ‘content’ and ‘rhetorical organisation’, however both groups rated ‘language use’ similarly and both distinguished between writing ability and language proficiency in their evaluations of the essays. Cumming also found differences between the groups in terms of their decision making behaviours. The inexperienced raters did considerably more editing of errors while rating the essays, while the experienced raters commented more on how they were distinguishing between the rating categories.

Vaughan (1991) used think-aloud protocols to investigate the behaviour of nine experienced raters using a six-point holistic assessment scale to assess six borderline ESL compositions. She identified several approaches to holistic assessment, such as the ‘grammar-oriented rater’, or the ‘first impression dominates’ approach in which raters make qualitative judgments about a text after having read the first couple of lines. Vaughan suggests that, while there was agreement amongst raters concerning many texts, when confronted with essays that did not clearly fit the holistic assessment guidelines, raters often resorted to their own assessment style to help them assign a score.

Weigle (1994) analysed protocols of four inexperienced raters scoring the same ESL compositions before and after training. The study showed that training helped raters understand the scoring criteria and modify their expectations of student writing. The protocols also showed that although training raised the awareness of the importance of inter-rater reliability and was successful in bringing extreme scores into an acceptable range, raters maintained their particular rating tendencies.

These studies indicate differences in the rating strategies employed by experienced versus inexperienced raters, and suggest some variability amongst individual raters. There is also some evidence that rater training sessions have a limited effect in changing raters’ inherent rating characteristics. This study was designed to contribute to this line of research.

### Methodology

#### Subjects

Six raters, all of whom were teachers with the NSW Adult Migrant English Service (AMES), participated in the research. They were naive raters in the sense that they had no previous experience with the access: test, either as raters or as item writers. Three raters were experienced teachers and three were inexperienced. The inexperienced teachers had one year or less adult ESL teaching experience. The experienced teachers had considerable adult ESL experience and substantial experience in the areas of literacy teaching, student assessment and placement, and curriculum design. The three experienced teachers had completed, or were enrolled in Masters degree programs of TESOL/Education. One inexperienced rater was enrolled in a Masters degree program of TESOL, one in a Graduate Diploma of TESOL, and one had completed a Graduate Certificate in TESOL.

#### Data

The two task types used for the ratings focused on quite different areas of functional purpose. One task, Text type 1, was a letter giving travel advice to a friend, and was typical of the tasks in Section 1 of the access: test, either as raters or as item writers. Three raters were experienced teachers and three were inexperienced. The inexperienced teachers had one year or less adult ESL teaching experience. The experienced teachers had considerable adult ESL experience and substantial experience in the areas of literacy teaching, student assessment and placement, and curriculum design. The three experienced teachers had completed, or were enrolled in Masters degree programs of TESOL/Education. One inexperienced rater was enrolled in a Masters degree program of TESOL, one in a Graduate Diploma of TESOL, and one had completed a Graduate Certificate in TESOL.

#### Procedures

Two protocol collection sessions were held six weeks apart. Text type 2 was rated in the first session and Text type 1 in the second. The six-week interval was designed to reduce any practice effect. The raters were asked to rate the texts
in the order they appeared in the booklet using a six-point rating scale. The scale was a numerical five-point scale where 0 = no assessable performance and 5 = competent writer of English as a Second Language, but there were no other descriptors. The raters were also asked to respond to the texts in whatever way they felt most comfortable, but at the same time were encouraged not to edit or vet their responses, but to verbalise all thoughts that occurred during the rating process. To ensure that the protocols were collected under identical conditions the raters followed exactly the same procedures on each occasion.

Coding categories

Raters’ comments were transcribed in full, with stress, pauses, repetitions, sighs, laughter, and questions included. The transcripts were claused and a coding system was developed on the basis of raters’ comments. Table 1 presents the fourteen categories used for the analysis of the protocols. These categories are similar, but not identical, to those established by Vaughan (1991). This table also gives an example of typical positive and negative comments made by the raters under each category. Once the coding system had been finalised, all the transcripts were coded for instantiations of comments relating to the above categories. The number of mentions for each category were then counted in each text, for each rater.

Results and discussion

In this section trends for the whole group and the sub-groups, experienced and inexperienced teachers, for both text types are presented and discussed in relation to frequency and thematic prominence. Frequency relates to those categories commented on most frequently; thematic prominence relates to those categories mentioned first by raters for each text. For each of these two features similarities and differences in rating patterns across task types and rating patterns according to rater profile are discussed.

Frequency

Table 2 presents the overall rankings for all categories for the two text types. The categories are ranked from 1–14 (1 = the most commented on category, 14 = the least commented on category). As can be seen, a consistent pattern emerged across all raters with the categories ‘organisation’, ‘grammar’, ‘task fulfilment’ and ‘mark’ being the most frequently mentioned for both text types. However, although the criteria ‘organisation’, ‘grammar’ and ‘task fulfilment’ appear to be salient to both text types, we cannot assume that when commenting on a particular trait, raters were referring to the same sorts of features in both text types. A comparison of the specific features commented on within any of the above traits revealed some variation between task types. A comparison of the specific features commented on within any of the above traits revealed some variation between task types.

Organisation

The high ranking of ‘organisation’ as a salient feature is in accordance with the findings of Deiderich et al. (1961), Freedman (1979) and Breland and Jones (1984). Comments categorised as ‘organisation’ in the protocols of the persuasive texts (Text type 2), included references to the introduction, whether the thesis was located in the first paragraph, whether a point of view was substantiated with supporting evidence, whether the arguments were organised logically and if the conclusion was appropriate.
‘Grammar’ emerged as being particularly salient for both text types and this tends to confirm the findings of Deiderich et al. (1961) and Harris (1977). All raters evaluated performance on a variety of grammatical features at the sentence or clause level for both task types. These typically included features such as the use of articles, word order, the use of conditionals and the ability to use varied and complex sentence structures. However, some differences emerged in the specific grammatical features some of the raters attended to — or said they would expect to find — in the different text types. When assessing the persuasive texts, some raters praised writers who were able to use grammatical resources to construct a generalised argument while in rating the interpersonal texts some raters specifically mentioned grammatical features which helped create the appropriate tenor, and that generally this meant the text had a more informal spoken feel about it. Other grammatical features mentioned included the use of contracted grammatical forms, informal lexis, and appropriate use of modality. The following comments illustrate this variation between the task types.

- **Grammar in persuasive texts:**
  
  `um ... problems with generalisation in line 7 'computer is very difficult to buy' so it should be a computer is or computers are so um I don't know ... generalisation I think is part of this kind of argument text and ah grammatically it's not handled well (Rater 1)`

  `this school should prepare' ... it should really be the school ... they haven't sort of like used the generic term the school or school (Rater 2)`

  `um the writer is in fact able to structure an argument ah to present an argument um without using 'I' so drawing on other sorts of interper- sonal resources um and that's quite effective (Rater 3)`

- **Grammar in interpersonal texts:**
  
  `postscript is quite effective just gives a little bit more um family details and it's quite reasonable to put those in a postscript (Rater 2)`

  `paragraphs, I mean fairly well organised in terms of requirements, like X in the first part Y in paragraph two (Rater 5)`

### Table 2: Overall ranking of categories in order of importance

<table>
<thead>
<tr>
<th>Overall ranking</th>
<th>Persuasive text</th>
<th>Interpersonal text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall performance</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>Task fulfilment</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Meaning</td>
<td>14</td>
<td>10</td>
</tr>
<tr>
<td>Appropriacy</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>Cohesion</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td>Organisation</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Spelling</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>Grammar</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Punctuation</td>
<td>13</td>
<td>9</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Layout</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Script</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Length</td>
<td>11</td>
<td>13</td>
</tr>
<tr>
<td>Mark</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

1 = most frequently mentioned category

14 = least mentioned category

For Text type 1 (the interpersonal texts), raters referred to features such as the opening and closing of the letter, whether the advice was presented logically in paragraphs and whether a postscript, if included, was appropriate. The following extracts are illustrative of comments on ‘organisation’ from both text types.

- **Organisation in persuasive texts:**
  
  `and then after the conclusion which should be the last bit there’s another thought tacked on, ah, which should have been in the body of the argument (Rater 1)`

  `the opening paragraph was well written ... but um a good introduction expressed a view point but then it seemed to get into a bit of a ramble (Rater 6)`

- **Organisation in interpersonal texts:**
  
  `postscript is quite effective just gives a little bit more um family details and it's quite reasonable to put those in a postscript (Rater 2)`

  `paragraphs, I mean fairly well organised in terms of requirements, like X in the first part Y in paragraph two (Rater 5)`

### Grammar

`Grammar’ emerged as being particularly salient for both text types and this tends to confirm the findings of Deiderich et al. (1961) and Harris (1977). All raters evaluated performance on a variety of grammatical features at the sentence or clause level for both task types. These typically included features such as the use of articles, word order, the use of conditionals and the ability to use varied and complex sentence structures. However, some differences emerged in the specific grammatical features some of the raters attended to — or said they would expect to find — in the different text types. When assessing the persuasive texts, some raters praised writers who were able to use grammatical resources to construct a generalised argument while in rating the interpersonal texts some raters specifically mentioned grammatical features which helped create the appropriate tenor, and that generally this meant the text had a more informal spoken feel about it. Other grammatical features mentioned included the use of contracted grammatical forms, informal lexis, and appropriate use of modality. The following comments illustrate this variation between the task types.

- **Grammar in persuasive texts:**
  
  `um ... problems with generalisation in line 7 'computer is very difficult to buy' so it should be a computer is or computers are so um I don't know ... generalisation I think is part of this kind of argument text and ah grammatically it's not handled well (Rater 1)`

  `this school should prepare' ... it should really be the school ... they haven't sort of like used the generic term the school or school (Rater 2)`

  `um the writer is in fact able to structure an argument ah to present an argument um without using 'I' so drawing on other sorts of interpersonal resources um and that's quite effective (Rater 3)`

- **Grammar in interpersonal texts:**
  
  `I'm very glad to know that you're coming to see us' so um you know friendly relationship is established with the um ... reader of the letter um ... the language is um, you know, quite casual 'you don't have to worry' and contractions and so on um (Rater 1)`

  `the person seems to have quite a ... degree of control in the sense ... that they can um go into a rather an informal mode in the first paragraph ... um 'take your favourite jeans and t-shirts and come' now okay um (Rater 2)`

  `um very um spoken like appropriate to the text things like 'so don't worry about the food' (Rater 3)`
‘Mark’ is a category which was mentioned often by both experienced and inexperienced teachers alike and signified the overall value a writer assigned to a text. The protocols indicated that generally there was a certain amount of unease amongst raters about assigning a mark in the absence of specific criteria against which to evaluate texts. As seen in the following comment by Rater 3:

its getting harder to rate because ... um the more I read um ... it just feels like it's not that it was a it was really clear to start with but um there are more things that I'm trying to think about now and I guess I'm also trying to compare the text as I read them with the text before knowing that by assigning a number to a particular text that I'm that I am effectively fleshing out what those numbers mean

This unease may have been an indication of personal differences in rating style, but nonetheless several raters expressed the dilemma that they did not know what the characteristics of a 3 or a 4 were, or that they would like to be able to give half marks.

Some raters progressively tried to establish some sort of framework for themselves within which to carry out the assessments. For two raters this meant attempting to norm-reference the texts as they went along and in so doing they expressed concern that perhaps they were being overly harsh or lenient, in which case all their scores would need to be scaled up or down to arrive at ‘true’ scores. As Rater 1 commented:

I think that how one rates these things might well depend on the order in which they are given to us ... because you're comparing it just with the one that's gone before ... the few that have gone before

Other raters attempted to establish criteria for different points on the scale. For example, Rater 3 in the following comment defines the significance of assigning a mark of three.

I'm remembering when I did the previous um session that I had in my mind a notion that when I said something was a three it meant that there was a sense of um of generic coherence about the text and that students had basically come to grips with putting a clause together even though there was some sort of ... errors still at the clause level
This attempt to establish criteria may indicate that some teachers in this study were accustomed to assessing written texts with reference to explicit performance criteria or assessment checklists. There were some indications from the transcripts that the lack of specific assessment criteria made the exercise a frustrating one for several of the teachers, particularly the experienced ones.

**Differences in rating patterns across task types**

Table 2 indicates that the fourth ranked category differed for both task types. For the persuasive text, the category ‘cohesion’ emerged as a salient trait, while for the interpersonal text it was ‘appropriacy’. This is discussed below.

**Cohesion**

‘Cohesion’ appeared to be salient for the persuasive texts but not the interpersonal texts. Some raters equated the sophistication of a written persuasive text with the writer’s ability to use cohesive devices effectively. This is not surprising because these devices contribute to overall task fulfilment, that is, they aid in the construction of an argument that is essentially sequential and logical in nature (Halliday 1985). The following examples provide a flavour of the comments typically made by experienced and inexperienced raters for persuasive texts:

- **Cohesion in persuasive texts:**
  - “the logical connections are very good here and the way he shows us a path through the argument and so ‘it is my firm conviction’ …” (Rater 1)
  - “connectives is pretty good um use of references …” (Rater 4)
  - “um has … has a good understanding of use ah references you know how using the references ah to refer back to point back to the … to point back in the article” (Rater 6)

The results as recorded in Table 2 suggest that cohesion was less salient for the interpersonal texts. Nonetheless, an analysis of these texts revealed that the more competent writers did use a considerable number of causal conjunctions to signpost the move from description to advice giving, although none of the raters commented on this. The reason may lie in the fact that this text type is closer to spoken language and conjunctive links such as those used in speech (for example, ‘so’, ‘but’, ‘then’) are less salient than the types of links used in persuasive texts.

**Appropriacy**

‘Appropriacy’ appeared to be particularly salient to the interpersonal texts, but less so for the persuasive texts. This may be because the interpersonal text (a letter to a friend) is closer to the spoken end of the spoken-written mode continuum (Martin 1984). As such we might expect the text to display features of informality and friendliness; that is, to be more spoken-like. Raters praised writers who succeeded in creating a ‘friendly’ tone, while those texts which sounded too distant or formal were seen as inappropriate; a text might have provided adequate advice and suggestions, but if it was written in a formal, impersonal style raters felt this seriously undermined the overall effectiveness of the piece of writing. In some cases, raters made specific mention of grammatical or lexical choices which were appropriate to a personal letter and which created the appropriate tenor. Sometimes, a combination of features contributed to the overall appropriacy (or inappropriacy) of the text. The following extracts are illustrative of these interpretations.

- **Appropriacy in interpersonal texts:**
  - “the tone is peremptory rather than friendly” (Rater 1)
  - “a rather a friendly and pleasant overall mood in the thing quite a nice sort of tenor” (Rater 2)
  - “it’s quite important that the tenor’s okay” (Rater 3)
  - “I’ll catch up later again by phone … with love” very appropriate concluding remarks (Rater 4)

Features such as those mentioned by the raters in the examples above distinguish interpersonal from expository texts, which are closer to the written end of the spoken-written mode continuum. Expository texts, on the other hand, often employ a neutral tenor to achieve their functional purpose of being impartial and rational (Martin 1985).

There was some discrepancy between raters, however, in respect to the type of written response they considered appropriate for the particular text task. Some raters took the view that the text was written to be read — that it was a formal written argument/submission to a committee; that is, an analytic exposition (Martin 1985). These raters penalised writers whose texts sounded too spoken-
like, seeing it as a sign of a lack of appropriacy and sophistication. The following extracts are illustrative of this view.

- Appropriacy in persuasive texts:
  *so I mean the whole spoken nature of it devalues it as a piece of writing, I think* (Rater 1)  
  *and ah the language lacks a sophistication of of some of the other written text we’ve seen because it reads more like a spoken text* (Rater 1)  
  *this one is very much a spoken sort of folksy kind of thing* (Rater 2)  
  *un it’s a very spoken like text and in fact the student concludes by saying ‘thanks for listening’* (Rater 3)

Other raters, however, were prepared to accept texts which were more formal in style and could have been written to be read aloud at a public forum; that is, hortatory texts (Martin 1985). These raters believed that the ability of the writers to mobilise the grammatical and lexical resources to produce such a text showed a high degree of sophistication. The following extracts illustrate this belief.

- Appropriacy in persuasive texts:
  *so it starts off ‘ladies and gentlemen’, but it doesn’t sign at anything, so it is obviously, um may be, it could be a spoken appeal* (Rater 2)  
  *it comes across as almost a a you know quite a sophisticated … speech style presentation um you know the opening of ‘ladies and gentlemen’ and at the end, oh where is it, ‘I therefore urge you to’ it’s it’s following a a definite … campaign … ab campaign style* (Rater 5)

Overall, the varying patterns in the ratings across task types suggests that cohesion is a criterion that is salient to the rating of persuasive texts, while appropriacy is salient to the rating of interpersonal texts.

**Rating patterns according to rater profile**

Individual rater trends are summarised in Tables 3, 4 and 5. Table 3 presents the total number of comments made by each rater at each session and the categories mentioned most frequently by raters for both text types. Tables 4 and 5 present the categories ranked in order 1–14, (1= most talked about category, 14= least talked about category, * = category not mentioned by this rater) for each rater for each text type. Table 4 summarises the rankings for the persuasive texts and Table 5 presents rankings for the interpersonal texts.

### Table 3: Most commented on category (frequency)

<table>
<thead>
<tr>
<th>Rater</th>
<th>Persuasive text</th>
<th>Interpersonal text</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total number of</td>
<td>Total number of</td>
</tr>
<tr>
<td></td>
<td>comments</td>
<td>comments</td>
</tr>
<tr>
<td></td>
<td>1st most frequently commented upon</td>
<td>1st most frequently commented upon</td>
</tr>
<tr>
<td></td>
<td>category (as % of</td>
<td>category (as % of</td>
</tr>
<tr>
<td></td>
<td>total comments)</td>
<td>total comments)</td>
</tr>
<tr>
<td>Rater 1</td>
<td>241 G21</td>
<td>252 G30</td>
</tr>
<tr>
<td>Rater 2</td>
<td>125 O24</td>
<td>221 G22</td>
</tr>
<tr>
<td>Rater 3</td>
<td>152 O34</td>
<td>111 O18</td>
</tr>
<tr>
<td>Rater 4</td>
<td>102 O27</td>
<td>116 O24</td>
</tr>
<tr>
<td>Rater 5</td>
<td>141 G23</td>
<td>186 G33</td>
</tr>
<tr>
<td>Rater 6</td>
<td>85 O22</td>
<td>152 T20</td>
</tr>
</tbody>
</table>

O = Organisation  
G = Grammar  
T = Task fulfilment

Raters 1–3 experienced adult ESL teachers; Raters 4–6 inexperienced adult ESL teachers

**Persuasive texts**

Some clear patterns emerged from the data for both groups of teachers rating the persuasive texts. As Table 3 shows, ‘organisation’ and ‘grammar’ were the most commented upon categories for both groups of teachers. However, the high ranking given to ‘cohesion’ (as indicated in Table 2) is shown in Table 4 to be largely due to comments by experienced teachers. While Rater 1 mentioned cohesion often, and Rater 2 also made a fair number of comments, cohesion was less salient for Rater 3 who was more like the inexperienced teachers on this category.

In general, the inexperienced teachers commented less on ‘cohesion’. However, the rankings for Rater 4 appear to correlate well with the rankings of the experienced teachers. While Rater 1 mentioned cohesion often, and Rater 2 also made a fair number of comments, cohesion was less salient for Rater 3 who was more like the inexperienced teachers on this category.
the development of their argument was well written and um I don’t think that there are that many problems with this student.

The data suggest general agreement between both groups of teachers with the categories ‘organisation’, ‘grammar’ and ‘task fulfilment’ being salient for both groups. However, the data also suggest that the number of comments regarding cohesion made by the experienced raters was responsible for raising the ranking of this category in Table 2.
‘grammar’ were mentioned frequently by this group, but the pattern of rankings was less marked. The number of comments made by the inexperienced teachers regarding ‘task fulfilment’ influenced the ranking of this category in Table 2.

Only Rater 4 of the inexperienced teachers commented frequently on ‘appropriacy’, as did the experienced teachers. In fact Rater 4 followed a very similar rating pattern to the experienced teachers with the interpersonal texts. Raters 5 and 6, on the other hand, rarely mentioned ‘appropriacy’. Their scores for these texts, however, did not differ markedly from the other raters so clearly these teachers could determine when a writer had fulfilled the requirements of the task, which presumably included establishing an appropriate tone. Their inexperience may have contributed to less explicit mention of the features which contributed to this effect.

For this task type, there was similarity between the ranking of the first three categories mentioned most frequently for both groups of teachers. A difference between the two groups was that the experienced teachers had a broader range of responses to the texts and referred to more precise linguistic features in the texts. The inexperienced teachers were, in general, less explicit although one of the inexperienced teacher’s rankings was very similar to the experienced teachers.

Thematic prominence

The overall trends for the whole group of raters in the categories given thematic prominence (ie those mentioned first) by raters are presented in Table 6.

<table>
<thead>
<tr>
<th>Feature</th>
<th>% of times feature mentioned first</th>
<th>Feature</th>
<th>% of times feature mentioned first</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mark</td>
<td>25</td>
<td>Layout</td>
<td>26</td>
</tr>
<tr>
<td>Layout</td>
<td>21</td>
<td>Overall performance</td>
<td>20</td>
</tr>
<tr>
<td>Overall performance</td>
<td>13</td>
<td>Mark</td>
<td>18</td>
</tr>
<tr>
<td>Task fulfilment</td>
<td>10</td>
<td>Length</td>
<td>10</td>
</tr>
<tr>
<td>Length</td>
<td>10</td>
<td>Appropriacy</td>
<td>10</td>
</tr>
<tr>
<td>Organisation</td>
<td>8</td>
<td>Organisation</td>
<td>8</td>
</tr>
<tr>
<td>Appropriacy</td>
<td>8</td>
<td>Task fulfilment</td>
<td>4</td>
</tr>
<tr>
<td>Script</td>
<td>4</td>
<td>Spelling</td>
<td>4</td>
</tr>
</tbody>
</table>

Similarities in rating patterns across task types

Overall, the data suggest that there was a high degree of similarity between the categories raters mentioned first across the two text types. As Table 6 indicates, features consistently given thematic prominence for both text types related to ‘layout’ and ‘overall performance’. In many cases a ‘mark’ was assigned after an initial reading of the text. The fact that raters most often began the assessment of a text with an evaluative comment on ‘overall performance’ or ‘layout’, or the assignment of a ‘mark’ suggests that initially raters are oriented to the features of texts that strike them visually or impressionistically and are oriented to a global assessment. In this respect the results agree with the findings of Stewart and Grobe (1979) and Charney (1984).

Overall percentages indicate that many of the categories given thematic prominence (eg layout, length, overall performance, script and spelling) did not correspond with the categories most frequently commented on. Certain features — ‘grammar’, ‘vocabulary’ and ‘punctuation’ — were never mentioned first. The fact that none of the raters began rating a text with comments on ‘grammar’ or ‘vocabulary’ suggests that these features do not lend themselves to impressionistic judgment and that raters may be more considered in their assessment of these criteria.

The quality of features such as layout, handwriting, spelling or text length may predispose a rater either positively or negatively to a text before a more detailed assessment is attempted. In the following extracts Rater 4 was initially impressed with the layout of a persuasive text, but by the end of the rating had severely penalised the writer because the argument was not clear.

<table>
<thead>
<tr>
<th>Feature</th>
<th>% of times feature mentioned first</th>
<th>Feature</th>
<th>% of times feature mentioned first</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mark</td>
<td>25</td>
<td>Layout</td>
<td>26</td>
</tr>
<tr>
<td>Layout</td>
<td>21</td>
<td>Overall performance</td>
<td>20</td>
</tr>
<tr>
<td>Overall performance</td>
<td>13</td>
<td>Mark</td>
<td>18</td>
</tr>
<tr>
<td>Task fulfilment</td>
<td>10</td>
<td>Length</td>
<td>10</td>
</tr>
<tr>
<td>Length</td>
<td>10</td>
<td>Appropriacy</td>
<td>10</td>
</tr>
<tr>
<td>Organisation</td>
<td>8</td>
<td>Organisation</td>
<td>8</td>
</tr>
<tr>
<td>Appropriacy</td>
<td>8</td>
<td>Task fulfilment</td>
<td>4</td>
</tr>
<tr>
<td>Script</td>
<td>4</td>
<td>Spelling</td>
<td>4</td>
</tr>
</tbody>
</table>

Very impressive piece of work ah in front of me um
I think this person has failed … to sort of put forward clear arguments for them in … the form of for or against the topic … ah this this person has sort of confused me no … um I’ll give him a two

If raters are influenced by these superficial features, it may be advisable to include a criterion such as ‘surface features’ to enable raters to make an overall evaluation of these traits. Otherwise there is the danger of the halo effect where the overall qualities of a text may influence the ratings of other criteria. McColly (1970: 151) points out:

…but when a judge thought he was judging some specific quality he was actually responding to his general impression or to the overall quality of the essay.
Once the salient criteria for particular text types have been determined, the next step in the development of either a multiple trait scoring instrument, or the scoring rubric for a holistic scale, involves decisions about the use of performance descriptors. On the one hand, some test developers (Davies: 1992; McDowell: 1995) believe that detailed descriptors should be avoided because they can actually disadvantage some candidates. Davies (1992: 14) states:

[T]he paradox is that through the attempt to refine proficiency scales by removing their defects ... the precisioning of the descriptors tends more and more towards a list or bank of test items. Descriptors which are usable in an objective sense are test items.

In other words, raters may seek out features mentioned in the descriptors and if they do not find them, or overlook them, they may penalise candidates. However, where performance descriptors are felt to be useful for raters, test developers have two options. They can develop task specific descriptors for each text type, but this would be time consuming and expensive. It also places a heavy cognitive load on raters asking them to juggle different scales for several text types in the same rating session. The alternative is the development of generic descriptors for the salient criteria at a range of levels. The results of this investigation, suggest that generic descriptors should be written in terms such as 'appropriate grammar' or 'appropriate organisation' for a given context, as in the access: rating scale.

Implications for rater training

The use of generic descriptors puts the onus on rater trainers to ensure that all raters are in agreement as to what is meant by ‘appropriate’ grammar, organisation and task fulfilment for a particular context. This implies that rater trainers have analysed the task types which appear in a given test paper so that they are aware of the features that would contribute to a text fulfilling its social purpose. These broad features, without being too prescriptive, then need to be communicated to raters.

Task analysis may also reveal that for any given prompt there may be several acceptable responses so that raters must be trained not to interpret the prompt too narrowly. This study revealed that in the absence of training prior to the rating sessions some raters adopted a very narrow view of what constituted a written argument and penalised writers who had written texts which had a more spoken flavour about them.
Notes

1. WWB Stylefiles numerically reports factors such as sentence and word length, sentence type, sentence openings, word class count, total number of words and sentences and the percentage of word types in each prose sample (see McDonald et al. 1982).

2. Predecessor of IELTS.

References


Notes on contributors

Geoff Brindley
Geoff Brindley is a Senior Lecturer in the School of English and Linguistics at Macquarie University and Research Coordinator in the National Centre for English Language Teaching and Research (NCELTR). He is the author of a range of publications on language proficiency assessment, second language acquisition and curriculum design.

Susan Delaruelle
Susan Delaruelle has worked with NSW AMES since 1981, recently specialising in the field of assessment of both proficiency and achievement. She worked on the development of the access: and step: tests which involved item writer training, task design, trialling test materials, and assessor training and moderation workshops. More recently, she designed assessment tasks for the Certificate in Spoken and Written English (CSWE) and co-authored Assessment and moderation: A professional development resource for teachers and trainers in adult ESL (NCELTR 1996).

Lesleyanne Hawthorne
Lesleyanne Hawthorne is a Senior Lecturer in the Centre for Cultural Studies in Health at the University of Melbourne, and has 20 years experience in both research and direct service delivery with NESB migrants and international students. Lesleyanne is the author of Refugee: The Vietnamese experience (OUP 1982); Making it in Australia (Edward Arnold 1988); and Labour market barriers for immigrant engineers in Australia (AGPS 1994), as well as co-author of Immigrants and the professions in Australia (1997) and English for occupational purposes (1988).

Kathryn Hill
Kathryn Hill has a BA (Indonesian, Arabic and Politics), a Diploma in Education (Indonesian and ESL) and an MA (Applied Linguistics), all from the University of Melbourne. Since 1992, she has been working as Research Fellow at the Language Testing Research Centre at the University of Melbourne. Her recent interests include the use of test-taker feedback, rater characteristics, teacher proficiency testing and the study of LOTE in primary schools.
Susan Hood
Susan Hood is a Principal Lecturer in the English Department at the Hong Kong Polytechnic University. Prior to taking up this post, she worked for some years as senior curriculum coordinator for the NSW Adult Migrant English Service and more recently as a lecturer in applied linguistics and TESOL at the University of NSW. She was the coordinator of the development of the reading and writing components of the access: test.

John Langille
John Langille is a teacher of English with the Higher Colleges of Technology, United Arab Emirates. He was previously a Senior Research Associate at NCELTR where he was involved in a variety of test development and evaluation projects.

Brian Lynch
Brian Lynch is Associate Professor and Director of the Centre for Communication Skills and ESL at the University of Melbourne. He has worked on English language programs in China, Mexico, Armenia and the United States, and has published in the areas of program evaluation and language testing.

Tim McNamara
Tim McNamara is Associate Professor in Applied Linguistics at the University of Melbourne, where he is also Senior Associate and former Director of the Language Australia Language Testing Research Centre. He has published widely on language testing, including Measuring second language performance (Addison-Wesley Longman 1996). He is working on a new book on language testing, and is part of the TOEFL 2000 project based at ETS in Princeton.

Colin McNaught
Colin McNaught is the General Manager (Operations) at NCELTR. He has been involved in TESOL in teaching, teacher training and management roles in Australia and overseas since 1977. He joined NCELTR in 1992 as Project Manager for the access: language testing project and was responsible for coordinating the test development, administration and marking arrangements for that project and for establishing test centres overseas. During the same period he also helped establish the test administration arrangements for the step: test for onshore temporary residents.

Janne Morton
Janne Morton works in the Centre for Communication Skills and ESL, University of Melbourne. Prior to this, she worked in the Language Testing Research Centre, University of Melbourne as a project officer on the access: test. Her research interests include discipline-specific discourse practices, writing pedagogy, and language testing. She has also recently developed an interest in child language acquisition.

Kieran O’Loughlin
Kieran O’Loughlin has worked in the ESL field for more than 15 years. He coordinated the initial trialling of the speaking component of the access: test at the NLLIA Language Testing Research Centre, University of Melbourne in 1992. His publications include a number of articles on language testing. Kieran is currently employed as a lecturer in ESL at the Centre for Communication Skills and ESL at the University of Melbourne.

Steven Ross
Steven Ross is professor of policy studies at Kwansei Gakuin in Kobe, Japan. He completed his PhD in second language acquisition at the University of Hawaii. He has been a senior research associate at NCELTR during 1994, and currently researches language assessment.

Gillian Wigglesworth
Gillian Wigglesworth is a lecturer in linguistics at Macquarie University. She has worked on a range of research projects in language assessment and second language acquisition and also has research interests in first language development. She has a number of publications in the areas of language assessment, second language acquisition and first language acquisition.

Donna Williams
Donna Williams has 14 years’ experience in TESOL at secondary and adult levels. She has also taught adult literacy and numeracy. From 1993 to 1995 she worked on the access: test as a rater, project officer and researcher. She now tutors in ESL at the University of Melbourne and edits publications.