

RESEARCH SERIES 11 2000

Studies in immigrant English language assessment

Volume 1

Edited by Geoff Brindley

RESEARCH SERIES **11** 2000



Studies in immigrant English language assessment

Volume I

Edited by Geoff Brindley

RESEARCH SERIES 11 2000



National Centre for English Language Teaching and Research
Macquarie University Sydney NSW 2109

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

Studies in immigrant English language assessment. Vol 1.

ISBN 1 86408 547 9

ISSN 1035 6487

1. Adult Migrant Education Program (Australia) – Evaluation. 2. English language – Study and teaching – Australia – Foreign speakers. 3. English language – Ability testing – Australia – Evaluation. 4. English language – Study and teaching – Political aspects – Australia. I. Brindley, G. P. (Geoff P.). II. National Centre for English Language Teaching and Research (Australia). (Series: Research series (National Centre for English Language Teaching and Research (Australia))); 11).

428.00715



© Macquarie University 2000

The National Centre for English Language Teaching and Research (NCELTR) is a Commonwealth Government Key Centre of Research and Teaching established at Macquarie University in 1988. The National Centre forms part of the Linguistics discipline at Macquarie University. NCELTR's Key Centre activities are funded by the Commonwealth Department of Immigration and Multicultural Affairs.

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.

The publishers have used their best efforts to contact all copyright holders for permission to reproduce artwork and text extracts and wish to acknowledge the following for providing copyright permission.

Appendix 1 pp 217–243 is reprinted from *Australian Second Language Proficiency Ratings* by D E Wylie and E Ingram, published by AGPS 1984. Commonwealth of Australia copyright reproduced by permission.

Figure 2 on page 22 reprinted from *Certificate II in Spoken and Written English*, NSW Australian Migrant English Services 1998, reproduced by permission.

Figure 4 on page 167 reprinted from *Certificate in Spoken and Written English II*, NSW Australian Migrant English Services 1995 reproduced by permission.

The reading assessment tasks in Appendix 3 on pages 249–262 are reprinted from the *Certificate in Spoken and Written English III*, NSW Australian Migrant English Services 1995 reproduced by permission.

The OTEN Information sheet reproduced in Appendix 3 on page 257 by permission of the Open Training and Education Network as part of the *Certificate in Spoken and Written English III (Further Study) Assessment*.

The letter 'Request to attend a full interview' reproduced in Appendix 2 on page 247 from The Commonwealth Employment Service, Commonwealth of Australia copyright reproduced by permission.

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 DTP: Helen Lavery and Vivien Valk

Cover design: Helen Lavery

Printed by: Centatime, Rosebery NSW

Contents

Figures	v
Tables	v
Acronyms	viii
Preface	ix
Chapter 1	
Assessment in the Adult Migrant English Program <i>Geoff Brindley</i>	1
Chapter 2	
Comparing AMEP assessments: A content analysis of three reading assessment procedures <i>Geoff Brindley</i>	45
Chapter 3	
Issues in the development of oral tasks for competency-based assessments of second language performance <i>Gillian Wigglesworth</i>	81
Chapter 4	
Task difficulty and task generalisability in competency-based writing assessment <i>Geoff Brindley</i>	125
Chapter 5	
Rater judgments in the direct assessment of competency-based second language writing ability <i>David Smith</i>	159
Chapter 6	
Individual differences and learning outcomes in the Certificates in Spoken and Written English <i>Steven Ross</i>	191

Appendixes

Appendix 1: ASLPR Scales	215
Appendix 2: ASLPR Reading Tasks	243
Appendix 3: CSWE Reading Assessment Tasks	247
Appendix 4: Examples of tasks used (manipulations not included)	261
Appendix 5: CSWE Writing Assessment Tasks	265

Lists of Figures and Tables

Figures

Figure 1: access: proficiency profile	8
Figure 2: Example of CSWE competency	22
Figure 3: Allocation of variables by task and task type	87
Figure 4: CSWE Writing Competency 14	167
Figure 5: ARMS data categories	194
Figure 6: CSWE Certificate I recursive path model of competencies achieved	196
Figure 7: Structural Equation Model: Certificate I	200
Figure 8: Structural Equation Model: Certificate II	201
Figure 9: Structural Equation Model: Certificate III	202
Figure 10: PACE criteria	204
Figure 11: Slow (1), Average (2) and Fast (3) Group Scatterplot	206
Figure 12: Histogram of PACE predictions for the test set (odd half)	207

Tables

Table 1: access: reading test	53
Table 2: ASLPR reading tasks	54
Table 3: CSWE reading tasks	55
Table 4: General description of test content	56
Table 5: Passage-based analysis: Tasks. Rater percentage agreement	59
Table 6: Item-based analysis: Rater percentage agreement	60
Table 7: Generalisability coefficients: Total ratings	60
Table 8: Passage-based analysis: Tasks	63
Table 9: Item-based analysis	64

Table 10:	Salient differences between assessments	64	Table 24:	Variance components for D-study 1 (1 rater x 1 task x 6 items): Competency 12	138
Table 11:	Language skills tested: Total ratings for each assessment	66	Table 25:	Results of D-studies for Competency 12	139
Table 12:	Passage-based input: Tasks	67	Table 26a:	CSWE Writing competency 10: Correlation matrix	140
Table 13:	Task and subject assignment	88	Table 26b:	CSWE Writing competency 12: Correlation matrix	141
Table 14a:	Certificate II, Competency 5, raw scores and Rasch estimates	90	Table 27:	Competencies 10 and 12: Principal factor analysis	143
Table 14b:	Certificate II, Competency 5, student evaluations (%)	90	Table 28:	Rating categories: Generic rating scheme — Item measurement report	149
Table 15a:	Certificate II, Competency 6, raw scores and Rasch estimates	91	Table 29:	Rater consistency in text classification	169
Table 15b:	Certificate II, Competency 6, student evaluations (%)	92	Table 30:	Relationship between reading strategy and number of texts passed	172
Table 16a:	Certificate II, Competency 7, raw scores and Rasch estimates	93	Table 31:	Textual features commented on by raters extraneous to performance criteria	173
Table 16b:	Certificate II, Competency 5, student evaluations (%)	93	Table 32:	Relationship between extraneous textual features commented on and rater reading strategy	173
Table 17a:	Certificate III, Competency 5, raw scores and Rasch estimates	93	Table 33:	CSWE Certificate 1 cross-validation results	197
Table 17b:	Certificate III, Competency 5, student evaluations (%)	94	Table 34:	Structural Equation Model outcomes Certificates I–III	202
Table 18a:	Certificate III, Competency 6, raw scores and Rasch estimates	94	Table 35:	Discriminant function analysis summary	205
Table 18b:	Certificate III, Competency 6, student evaluations (%)	95	Table 36:	Odds ratios for Certificate I by language group	209
Table 19:	CSWE Writing: Competencies 10 and 12 Task Measurement Report	132	Table 37:	Odds ratios for Certificate II by language group	210
Table 20:	CSWE Writing: Competencies 10 and 12 Rater Measurement Report	133	Table 38:	Odds ratios for Certificate I by migration category	210
Table 21:	CSWE Writing: Competencies 10 and 12 Item Measurement Report	134	Table 39:	Odds ratios for Certificate II by migration category	211
Table 22:	Competency 10: Variance Components for D-study 1 (1 task x 1 rater x 4 items)	137	Table 40:	Summary of individual differences analyses	212
Table 23:	Dependability (<i>phi</i>) coefficients for D-studies: Competency 10	137			

Acronyms

access:	Australian Assessment of Communicative English Skills
ACTFL	American Council on the Teaching of Foreign Languages
ALTE	Association of Language Testers in Europe
AMEP	Adult Migrant English Program
AMES	Adult Migrant English Service
ARMS	AMEP Research Management System
ASLPR	Australian Second Language Proficiency Ratings
CLA	communicative language ability
CSWE	Certificates in Spoken and Written English
CTCS	Cambridge-TOEFL comparability study
DIMA	Department of Immigration and Multicultural Affairs
EAP	English for Academic Purposes
EEC	English Education charge
ESL	English as a Second Language
FCE	First Certificate in English
FE	Functional English
IELTS	International English Language Testing System
ILR	Interagency Language Roundtable
IRT	item response theory
ISLPR	International Second Language Proficiency Ratings
NCELTR	National Centre for English Language Teaching and Research
NNS	non-native speaker
NS	native speaker
step:	Special Test of English Proficiency
TOEFL	Test of English as a Foreign Language
UCLES	University of Cambridge Local Examinations Syndicate

Preface

This volume provides an important set of in-depth, authoritative empirical analyses on specific aspects of the assessment instruments and procedures used over the past decade in the Adult Migrant English Program (AMEP). For adult educators within Australia, the book raises some sobering questions about the validity and interpretations of language competency assessments, the compatibility of related tests with one another, as well as decisions routinely made about people's language abilities or achievement. At the same time, the book provides an informed, substantive research base on which to evaluate these matters, to appreciate the sophistication and enormous potential of the AMEP's information system, as well as to prepare for future studies and improvements in program policies. I expect the volume will have a distinctive impact on thinking about English competency assessment in adult education programs throughout Australia. Acknowledging the considerable advances made in this area within the AMEP, the book leads the way forward in steps to refine and make these assessments more consistent, valid and coherent.

For language educators internationally, the volume is welcome for its overview of issues, practices, and resources for language assessment in Australia as well as for its presentation of innovations in the research methods related to language testing. Although the issues in language testing research addressed are unique to the Australian context in the 1990s, they link directly with numerous topical themes and innovations in language testing, education, and public policies internationally. In this respect, three themes on which Brindley and his collaborators have focused are especially worth highlighting for their broader significance:

- 1 establishing and maintaining assessment standards;
- 2 interfaces between testing and educational policies; and
- 3 multiple approaches to validation research.

All of the chapters are concerned with key themes in evaluating the standards and practices for English language assessments within the AMEP. To what extent are the assessments doing what they intend to do? Are they fair to diverse populations? What aspects of them might be inconsistent and thus be improved or made more informative? In short, are the assessment instruments, and the uses people make of them, valid and relevant to the program policies

and the clients they serve, and are they capable of refinements? These fundamental concerns have featured in many major initiatives related to language policies and assessment around the world in recent years.

As with the AMEP, initiatives to standardise language testing have been a principal unifying factor in many large-scale language programs or policy frameworks related to them. Other notable examples are the Council of Europe's framework for assessing life-long language learning (eg North 1995), the Language Benchmark Assessments for adults settling in Canada, which was modelled in part on AMEP's Certificates in Spoken and Written English (eg Norton Peirce and Stewart 1997), the proficiency guidelines of the American Council on the Teaching of Foreign Languages (eg ACTFL 1986) or diverse initiatives to improve bilingual opportunities or recognition among Hispanic Americans in the United States (eg Valdes and Figueroa 1994). In more general terms, Alderson, Clapham and Wall (1995:235–260), Groot (1990) and McNamara (1998) have reviewed existing efforts to maintain high levels of standards in language assessment to serve educational purposes and public policies, and they have argued for continuing research to strive to monitor and maintain these.

As in the AMEP, concerns for validity, fairness and relevance are vital because of the heavy reliance on teacher-administered assessment and outcomes-oriented criteria (Brindley 1998). Importantly, if decisions are to be made about people's language abilities, learning needs, or achievements, then assessments must ensure these are consistent across the diverse contexts in which, and populations for which, the program is administered. The weight of evidence from the studies presented here, however, is that such standardisation is exceedingly difficult to achieve in practice, due to a variety of variables and limitations, many of which the present research elucidates. These findings make further, systematic inquiry and the utilisation of relevant resources all the more important to assure the requirements of consistency, fairness, and validity in these assessment practices, the information they provide, and the decisions that follow from them.

In this regard, a central feature of this book is its exemplification of the diverse ways in which these matters need to be investigated. The studies reported are remarkably varied in their approaches to research as well as the aspects of assessment they address. Perspectives range from the micro-level of test items and raters' decision making to the macro-level of long-term language learning

outcomes among whole populations. Research techniques include content analyses, quasi-experimental manipulations, discourse analyses, generalisability comparisons, Rasch analyses, factor analyses, think-aloud verbal reports, as well as structural equation modelling. The studies span the conventional domains of language competency — reading, writing, and speaking — extending as well (in Ross's concluding chapter) into analyses of the long-term societal impact of English language programs.

Together this diversity is highly complementary, displaying the range of relevant possibilities for inquiry in this field (cf Clapham and Corson 1997; Cumming and Berwick 1996; Kunnan 1998). Collectively, the volume asserts the value of studious research into language assessment while challenging assumptions about assessment practices with solid, empirical evidence, intelligent analyses, as well as useful recommendations. The distinct message is that validation is a multi-faceted, complex, long-term process. Extensive, systematic research using multiple methods of inquiry into diverse facets of language assessment and the impacts that follow from them is necessary to strive to establish that language assessments are doing what people want them to, and if they are not, to be able to know how to modify them accordingly.

Alister Cumming

Head, Modern Language Centre

Ontario Institute for Studies in Education

University of Toronto, Canada

References

- American Council on the Teaching of Foreign Languages (ACTFL) 1986. *ACTFL proficiency guidelines*. Hastings-on-Hudson, New York: ACTFL
- Alderson, J C, C Clapham and D Wall 1995. *Language test construction and evaluation*. Cambridge: Cambridge University Press
- Brindley, G 1998. 'Outcomes-based assessment and reporting in language learning programs: A review of the issues'. *Language Testing*, 15: 45–85
- Clapham, C (volume ed) and D Corson (series ed) 1997. *Language testing and assessment*, volume 7 of *Encyclopedia of language and education*. Dordrecht, Netherlands: Kluwer
- Cumming, A and R Berwick (eds) 1997. *Validation in language testing*. Clevedon, Avon: Multilingual Matters

- Groot, P 1990. 'Language testing in research and education: The need for standards'. In J H A L de Jong (ed). *Standardization in language testing*. AILA Review 7: 9–23
- Kunnan, A (ed) 1998. *Validation in language assessment*. Mahwah, NJ: Lawrence Erlbaum
- McNamara, T 1998. 'Policy and social considerations in language assessment'. *Annual Review of Applied Linguistics*, 18: 304–319
- North, B 1995. 'The development of a common framework scale of descriptors of language proficiency based on a theory of measurement'. *System* 23, 445–465
- Norton Peirce, B and G Stewart 1997. 'The development of the Canadian Language Benchmarks Assessment'. *TESL Canada Journal*, 14, 17–31
- Valdes, G and R Figueroa 1994. *Bilingualism and testing: A special case of bias*. Norwood, New Jersey: Ablex

1

Assessment in the Adult Migrant English Program

Geoff Brindley

Introduction

The present volume brings together a range of research studies into aspects of the procedures used to assess and report the English language proficiency and achievement of adult immigrants enrolled in the Adult Migrant English Program (AMEP) in Australia. With the exception of Chapter 5, all of the studies were conducted between 1995 and 1998 as part of the Special Projects Research Program conducted by the National Centre for English Language Teaching and Research (NCELTR) at Macquarie University and funded by the Commonwealth Department of Immigration and Multicultural Affairs (DIMA).

The aim of this introductory chapter is to provide the background and rationale for the research studies reported in this volume and to situate them in relation to broader assessment issues. Although the studies relate specifically to Australian adult ESL programs, many of the assessment issues and dilemmas which they raise are currently preoccupying educational systems throughout the world. For this reason it is hoped that the research reported here will be of relevance and utility to teachers, test developers, program administrators and policy makers working in a range of other language teaching contexts.

The first part of this chapter describes a number of tests and assessment procedures used in the AMEP. These are:

- the Australian Assessment of Communicative English Skills (**access:**) (Brindley and Wigglesworth 1997), a standardised proficiency test used for immigration selection;
- the Australian Second Language Proficiency Ratings (ASLPR) (Ingram and Wylie 1984), a proficiency rating scale used for assessment at entry and exit to the AMEP; and

- the assessment tasks used in conjunction with the Certificates in Spoken and Written English (CSWE) (NSW Adult Migrant English Service [AMES] 1998) used to assess achievement of specific language competencies.

The principal focus in this discussion will be on those assessments which are the most commonly used to make decisions on learner proficiency and achievement within the AMEP curriculum, that is, the ASLPR and CSWE, although assessment issues which are also relevant to testing for immigration selection will be canvassed briefly.

This chapter will describe the purposes of each of these assessments, their theoretical underpinnings, and some of the issues and problems associated with their development and use. Some similarities and differences between the assessments will also be highlighted and explored. Arising from this overview, a number of key issues and questions will be identified which provided the stimulus for the research reported in this volume.

Background

The Adult Migrant English Program

The AMEP provides a range of English language learning opportunities to recently arrived immigrants and refugees to Australia. It is funded by the Commonwealth Department of Immigration and Multicultural Affairs (DIMA). The AMEP is one of the largest government-funded second language programs in the world, with an annual budget in excess of \$A92 million in 1998. It is delivered in all states and territories by 29 provider organisations and employs over 500 teachers across Australia in more than 100 teaching locations, as well as being offered in distance learning and self-study mode.

English language assessment of adult immigrants

Assessment for immigration decisions

In 1992 an English language proficiency requirement was introduced by the Australian government for applicants for migration to Australia in certain categories. Between 1993 and 1997, the Australian Assessment of Communicative English Skills (*access*), a standardised test of speaking, listening, reading and writing specially developed by a consortium of Australian educational institutions, was used to assess language proficiency for immigration purposes (see Brindley and Wigglesworth 1997). This test was subsequently replaced by the

International English Language Testing System (IELTS) General Training module (UCLES/British Council/IDPEA 1998).

Another English language test which has been used in the context of immigration policy is the Special Test of English Proficiency (*step*). This test was developed in 1994 as a result of a decision by DIMA to include English language proficiency as a requirement for the granting of permanent residence under a special 'fast track' visa category created by the government to facilitate the processing of applications for refugee status in Australia, mainly immigrants who had arrived from the People's Republic of China after June 1989 (see Hawthorne 1996). Between 1994 and 1996 this test was administered to over 10 000 candidates but has now been discontinued and will not be considered here.

Assessment in the AMEP curriculum

At entry to the AMEP, learners' proficiency in speaking, listening, reading and writing is assessed through an oral interview using the Australian Second Language Proficiency Ratings (ASLPR) (Ingram and Wylie 1984), a general proficiency rating scale.¹ (The ASLPR is described in detail below.) The results of ASLPR assessments are used along with other information to place clients in appropriate learning arrangements. ASLPR ratings are also used for end of course assessments within the AMEP and for reporting learning outcomes to the federal government and other external agencies.

No standardised syllabuses are prescribed within the AMEP. However, since 1994 a competency-based curriculum framework, based around the CSWE (NSW AMES 1998), has been used by organisations which provide AMEP courses as the basis for curriculum planning and assessment. This framework consists of Certificates corresponding to four levels of English language ability and provides a statement of the learning outcomes in the form of language competencies in oral interaction, reading and writing. Task-based assessments of the CSWE competencies are administered by teachers at periodic points throughout the program and at the end of a course. On the basis of these assessments, learners may achieve a statement of attainment or be awarded a Certificate if they achieve the requisite number of competencies. Since 1996 the CSWE competencies have been used, in conjunction with ASLPR ratings, for reporting learners' language gains to external authorities. Competency outcomes are recorded via the AMEP data management system, ARMS. In order to facilitate the planning and monitoring of program delivery, a set of 'bench-

marks' based on analysis of aggregate competency outcomes since 1996 has been developed. These describe average outcomes in terms of competencies achieved according to proficiency level, learning pace and hours of instruction (see Ross, Chapter 6, this volume).

Using different assessments: Issues and problems

The coexistence of a number of different assessment and reporting procedures within the AMEP curriculum has created a number of difficulties for both teachers and administrators. First, since an ASLPR rating provides only information on 'general proficiency', a learner's status in relation to specific competencies cannot be ascertained at entry to the program. As a result, it becomes impossible to report reliable gains in terms of competencies for purposes of program reporting and evaluation.

A second related problem arises from the fact that a general proficiency scale such as the ASLPR is not sensitive enough to indicate progress on the part of learners who have achieved a limited number of specific competencies. This applies particularly in the case of learners at very low levels of proficiency who may have little education (Cope et al 1994). Many of these learners may in fact have made progress in terms of competency attainment but these gains may be too small to register on the ASLPR scale. This discrepancy has potentially serious consequences in an environment in which funding authorities are looking for clear evidence of gains in order to demonstrate the effectiveness of the Program and is one reason for the adoption of CSWE competencies as a basis for reporting learner outcomes.

A third contentious issue which affects placement and exit assessment is the question of the relationship between CSWE, ASLPR and the 'Functional English' (FE) threshold which is used to make decisions concerning learners' eligibility for English language instruction. The FE level has assumed a good deal of importance as a decision-making tool in the AMEP. From March 1993, newly arrived immigrants and refugees have been entitled to 510 hours of government-funded English language tuition, or the hours it takes to reach the FE level, whichever is achieved first. The AMEP handbook (Department of Immigration and Ethnic Affairs 1995) states that 'functional English proficiency equates to an assessment of at least 2 on the ASLPR scale for all four macro skills — speaking, listening, reading, writing. *In other words, if any macro skill scores less than 2 on the ASLPR scale, the client is eligible for DIEA-funded English language tuition* (emphasis in original).

Although these restrictions were subsequently relaxed, the FE level may determine eligibility for the Program and the amount of tuition provided as well the payment or non-payment of the English Education charge (EEC). In addition, at exit from the AMEP, eligibility for further education or training opportunities may also depend on whether an immigrant's language level meets the FE criterion. In this regard, a level of ASLPR 2 in all four skills has conventionally been required as the minimum language proficiency for entry to many vocational training institutions.

However, the question of the equivalence between ASLPR, CSWE and — until recently — **access:** levels is shrouded in confusion. Currently, CSWE levels are roughly equated to ASLPR ranges for purposes of class placement within the AMEP but these equivalences have never been empirically investigated and are therefore open to question. Moreover, there is some evidence to suggest that teachers perceive discrepancies between ASLPR and CSWE assessments (Bottomley et al 1994). Similar problems in relation to equivalences between **access:** and ASLPR levels emerged at the time when **access:** was used for immigration selection testing. Some cases were reported of recently arrived immigrants enrolling in the AMEP who had failed to achieve the *Functional* proficiency level in the **access:** test but who were assessed by teachers as above this level on the ASLPR, thus making them ineligible for the English language classes for which they had already paid.

Given the importance of the educational decisions concerning learners which are made on the basis of ASLPR ratings, CSWE competency achievement and — until recently — **access:** scores, it is important to investigate the relationships between these assessments. In this way it becomes possible not only to ascertain likely reasons for discrepancies in the way learners are classified but also to identify areas of needed improvement in the development and use of the tools that are used to elicit performance.

It should be pointed out in the context of this discussion that a score or level on one of these assessments cannot be directly converted to another, despite the desire on the part of educational authorities for a quick and easy conversion procedure. This is because the assessments in question do not meet the requirements for test equating — that is, they do not 'measure the same thing in the same way' (Linn 1994). Notwithstanding this lack of direct equivalence, similarities and differences between the content and structure of each assessment can be identified. At the same time it is also possible to investigate the extent to

which the tasks used in each assessment are providing valid and reliable information on language proficiency and achievement. Put simply, in the words of Weir (1993:7):

In both proficiency and achievement testing there is a need for a clear understanding of what test tasks are measuring and how well they are doing it.

AMEP assessment procedures

The following section provides a description of **access:**, the ASLPR and CSWE. Since the ASLPR and CSWE are the tools used for curriculum-related decisions such as placement, progress and exit assessment within the AMEP, the discussion will centre mainly on these two assessments. However, the results of research into the **access:** test which are relevant to the issues under consideration will also be outlined briefly. Each of the assessments is considered in turn in terms of its background and purpose, content, administration and use. Issues relating to the validity and reliability of each are canvassed and a number of problems and unanswered questions are identified. The discussion of the ASLPR is somewhat more detailed and critical since it draws on the substantial body of research literature on 'direct' methods of language assessment such as the ACTFL Guidelines which have been widely used in language learning programs for many years. In this context, it should be noted that many of the issues and problems that are identified in the discussion are not specific to the ASLPR but apply equally to other assessment tools that use oral interviews as a basis for eliciting samples of language performance.

The Australian Assessment of Communicative English Skills (**access:**)

Purpose

The Australian Assessment of Communicative English Skills (**access:**) is a task-based test of English language ability which until 1997 was administered in the country of origin of applicants for immigration to Australia. It aims to sample the kinds of language which future immigrants might encounter in a range of social and occupational contexts in Australia.

The results of the test are reported in the form of a Proficiency Profile describing 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 summarising in general terms the types of performances which could be expected from candidates at different levels (see Figure 1). At the time when the test formed part of the immigration selection procedures, a 'Vocational Proficiency' level (corresponding to Level 5) was required by professionally skilled applicants in order to obtain a visa. Those applicants whose language level was assessed as less than 'Functional Proficiency' (Level 4) were required to pay an English Education charge (EEC) to cover the cost of language tuition in Australia via the AMEP prior to the granting of a visa, though some categories were exempted from the charge.²

Structure and administration

The test contains reading, writing, and listening skills modules and a test of oral interaction in the form of a structured interview with a native speaker interlocutor. A semi-direct form of the oral interaction test suitable for delivery in a language laboratory was also developed. This version was pre-recorded on cassette and was designed to be parallel in content to the direct interviewer-based version (O'Loughlin 1997). It was intended for use in centres where native speaker interviewers were not readily available and where laboratory facilities existed.

Validity and reliability

The **access:** test has been the subject of a range of research studies aimed at investigating aspects of its validity, reliability and practicality. These studies have concerned *inter alia*, the equivalence of the direct and semi-direct forms of the oral interaction module (Wigglesworth and O'Loughlin 1993; O'Loughlin 1995, 1997); interviewer behaviour in the oral interaction module (Morton et al 1997); the relationship between listening skills and item difficulty (Brindley 1997b); the relationship of rating criteria to writing text types (Delaruelle 1997); test-taker perceptions of test validity (Hill 1997); and the generalisability of oral and written tasks (McNamara and Lynch 1997). It is not possible to summarise all of these studies here. However, results of some of these studies relevant to the issues under discussion here will be outlined briefly below.

Brindley (1997b) investigated the relationship between listening skills and item difficulty in the **access:** listening skills test. He found 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

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 contexts.

Level Five

You can **read** and understand a wide range of 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.

Level Four

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 well enough to handle everyday communication adequately, despite some errors.

Level Three

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 basic communication in everyday situations but you make a lot of errors.

Level Two

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.

Level One

This means that you cannot read, write, understand or say anything in English or that you know only a few common words and phrases.

Figure 1: access: Proficiency Profile

Source: *access: Issues in language test design and delivery*. Research Series 9 (eds Geoff Brindley and Gillian Wigglesworth 1997) p 36

items. This lack of agreement was attributed 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. Brindley questioned the appropriacy of asking native speakers to introspect on the difficulty of items for language learners and called for more research into the test taking processes and strategies used by the actual population for whom the test is designed.

O'Loughlin (1997) examined the comparability of the face-to-face version and tape-mediated versions of the **access**: oral interaction test. He found that the tape version appeared more 'test-like' to candidates and that the live version offered the potential to elicit a richer sample of language, although the quality and quantity of the language sample elicited appeared to depend on the skill of the interviewer. He also found a clear preference on the part of candidates for the live format. In his examination of scoring patterns in the test, O'Loughlin uncovered a substantial mismatch between the levels assigned to candidates on the two versions. An analysis of rating patterns revealed this discrepancy to be due to significant bias in some of the ratings on both test formats. O'Loughlin suggested that these rating irregularities could be attributed on the one hand to one rater's difficulties in interpreting rating criteria and on the other to the interlocutor's failure to elicit an adequate sample of language in the live test. He went on to identify a number of factors which contributed to measurement error in oral tests, including candidate familiarity with the format, interlocutor behaviour and rater bias and suggested ways in which these factors might be controlled. Because of the multiplicity of factors influencing the oral assessment process, O'Loughlin issued a plea for caution in the interpretation and use of oral test scores.

Morton, Wigglesworth and Williams (1997) examined differences between interviewers who conducted the **access**: oral interaction test. They found a tendency on the part of raters to compensate for interviewers they perceived as being poor, and that non-native interviewers were more frequently classified as poor. Their analysis of the discourse of the structured tasks in the test versus an unstructured task (the role play) revealed clear differences in interviewer behaviour, with good interviewers providing more opportunity for candidates to produce their best performance. The researchers found clear evidence that variability in interviewer behaviour could have a marked effect on the ratings given to candidates and argued that this provides a justification for the use of a highly structured test.

McNamara and Lynch (1997) used generalisability theory (G-theory) to investigate the effects of different combinations of tasks and raters on the reliability of the oral and writing modules of the **access**: test. Their results indicated that single ratings produced unacceptably large proportions of error variance. On the basis of this finding, they concluded (1997:211) 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.

A number of themes emerge from these research studies into **access**: which are also of relevance to the other assessment tools used in the AMEP and which will be further explored in later chapters in this volume.

The Australian Second Language Proficiency Ratings (ASLPR)

Background

The ASLPR (Ingram and Wylie 1984) was commissioned by the Joint Commonwealth-States Committee on the Adult Migrant Education Program in the late 1970s as part of a major upgrading of English language programs for immigrants. Ingram (1981:109) explains this situation thus:

... it became evident that some clear notion of the path of development of second language proficiency was required as a framework within which courses could be planned. This framework needed to take into account, not just of syntactic and lexical mastery, but also of the tasks that learners could carry out in the language, especially in so far as those tasks were relevant to their needs as residents of Australia.

Purposes

The ASLPR serves a variety of functions in the AMEP. Language proficiency level according to the ASLPR, along with other factors such as age, educational background, first language and previous language study, is used as a basis for determining eligibility for the AMEP and for class placement. In addition, ASLPR gains achieved by learners are reported to funding authorities for accountability purposes. As mentioned previously, ASLPR levels also serve to determine eligibility for entry into further AMEP classes, government-funded training programs or tertiary study. In addition to these uses in the AMEP,

Ingram (1990:53) states that the ASLPR has been used to assess language skills for vocational registration; to assess the legal responsibility of defendants in court; to specify minimum levels of foreign language proficiency for teachers; and to develop data on rates of gain in language programs. Another development in recent years has been the increasing use of the ASLPR in workplace contexts for employment selection as well as for certification of proficiency for legal purposes (McIntyre 1995).

Scale descriptors

The ASLPR describes language performance at nine defined proficiency levels (0, 0+, 1-, 1, 1+, 2, 3, 4, 5) along a continuum which ranges from inability to function in the target language (zero) to native-like proficiency (5). (A full copy of the 1984 version of the ASLPR is included in Appendix 1.) Three additional levels (2+, 3+ and 4+) are also used by raters but have not been specifically defined (Ingram 1990:46). Commenting on the level descriptors, Ingram (1981:112) notes that ‘except for 0 and 5, the defined points are neither absolute nor discrete, each definition existing in the context of the whole scale and in relation to adjacent definitions’.

Unlike some scales which define different aspects of language performance separately, the ASLPR subsumes a range of features of language performance (eg grammar, vocabulary, pronunciation etc in the case of speaking) into a single level description.

Rating procedures

Ratings are awarded on the basis of candidates’ performance in an oral interview conducted by one or more native speaker interlocutors. Raters are required to match the behaviour of the testee to the level description which most closely resembles the behaviour observed. Separate reading tasks are usually administered as part of the interview. Some interviewers use recorded material accompanied by questions to rate listening ability, while others base their listening ratings solely on the candidate’s performance in conversational interaction. The latter seems to be the most common practice (McIntyre 1995). Writing tasks are usually undertaken by candidates independently of the interview.

Assessment tasks and materials

No standardised assessment materials or recommended assessment tasks or activities are provided with the ASLPR. This is on the grounds that:

... because language is significantly situation-dependent, it is necessary to elicit language behaviour in situations that are relevant to the learner or at least to ensure that the learner's inability to perform in a situation is not the result of unfamiliarity with that situation rather than the level of language development he or she has attained.

(Ingram 1990:52)

Task and text selection are left up to the interviewer who decides which elicitation materials to use on the basis of the candidate's perceived proficiency level, needs and interests. These materials are devised during ASLPR training sessions. ASLPR 'kits' containing a collection of texts and tasks for use in interviews have also been developed in many teaching centres in the AMEP.

Thirty minutes are recommended for assessing speaking, listening and reading and an hour for writing, with shorter times allowable in the case of lower level learners (Wylie and Ingram 1992:1). McIntyre (1995) found that in practice interview times ranged from ten to thirty minutes, depending on the perceived level of the interviewee.

Elicitation techniques

The ASLPR interview is carried out through three stages: a settling down, brief exploratory period, followed by a more searching analytic period in which the learners are extended to reach their full potential and finishing with a brief period at a level comfortable for the learner. The interviewer elicits a sample of language performance from interviewees by employing a range of techniques, including conversation, questioning strategies, visual stimuli and topic development strategies. (Wylie and Ingram 1992; Manidis and Prescott 1994). Closed *yes/no* questions are likely to be used at the exploratory stage of the interview and open *wh-* questions at the analytical stage as the latter are considered more appropriate for eliciting maximum language and enable the interviewer to pay more attention to the rater role (Manidis and Prescott 1994:33).

Theoretical basis

The ASLPR interview represents an example of the type of performance test in which information on testees' ability is elicited 'directly' in a situation which closely resembles the target language use situation:

... proficiency statements are related to and made in terms of the learner's actual language behaviour. Thus, rather than just measure

knowledge, direct instruments describe different levels of proficiency in terms of the sort of communication tasks that the learner can carry out and how they are carried out. (Ingram 1984a:5)

According to Ingram (1984a:11) the ASLPR measures 'general proficiency' which 'refers to the ability to use language in ... everyday, non-specialist situations'. He makes a distinction between 'proficiency' and 'communicative competence', arguing that the latter depends on non-language factors such as intelligence, personality traits and general knowledge (Ingram 1984a:16) which are probably not appropriate for the language tester to measure.

Validity and reliability

Ingram (1984b, 1990) reports high levels of inter-rater and intra-rater reliability in formal trials of the ASLPR, with correlations between raters generally above .9. On the basis of these findings Ingram (1990:59) concludes that:

The ASLPR does seem able to make valid statements about learners' language proficiency and users of the ASLPR, whether native or non-native English speakers, do seem able to interpret and apply the scale reliably and to make valid and reliable assessments of learners' language proficiencies.

As far as the validity of the ASLPR is concerned, Ingram (1990:54) cites the widespread acceptance of the scale as an indicator of its face validity. He also reports high concurrent validity coefficients from a comparison with the Comprehensive English Language Test (CELT) (Harris and Palmer 1986) (op cit p 55). In relation to predictive validity, Ingram (1995:25) refers to a study by Kellett and Cumming (1995) who found a high correlation between ASLPR levels and student performance in vocational education courses.

No published studies of the construct validity of the ASLPR are available. However Ingram (1995:23) refers to an unpublished study by Lee (1995) who concluded that:

- 1 The ordinal nature of the ASLPR levels is established.
- 2 The nature of the four macroskills as sub-scales of the ASLPR is established.
- 3 The ASLPR scale and its sub-scales seem able to uncover an ESL proficiency developmental path of learners from diverse L1 backgrounds and age groups covering two years.

The ASLPR: Issues and problems

Over the years, 'direct' methods of assessment such as the ASLPR and American Council on the Teaching of Foreign Languages (ACTFL) oral interview have been subjected to a good deal of critical scrutiny. Many of the key issues and problems associated with the 'real life' approach have been canvassed extensively in the literature (see, for example Bachman 1990; Brindley 1994; McIntyre 1995; McNamara 1996; Shohamy 1996) and will not be reiterated in detail here. However, certain aspects of the scale itself and of the oral interview method which are relevant to the present chapter will be discussed below.

Lack of evidence for construct validity

According to Ingram (1984a:11) 'the ASLPR seeks to measure the underlying general proficiency rather than the fulfilment of an absolutely specified task in an absolutely specified situation'. However, the scale includes descriptions of specific contexts, topics and functions and contains no model of underlying performance factors. McNamara (1996:76) sees this as a contradiction:

We thus have the contradictory position of, on the one hand, a performance requirement (a focus on actual instances of use) but, on the other, a principled exclusion of underlying general performance capacities in evaluations of test performance.

McNamara (op cit p 79) argues that failure to include factors other than contextual ones in the assessment stands in the way of validation:

Ingram's belief in the automatic validity of his procedure amounts to no more than a broad claim about its face validity; the apparent naïveté of this position stifles investigation into the hard questions about fairness and validity that need to be asked about this and kindred procedures. Of course Ingram is not alone in this complacency, which is an unfortunate characteristic of much practice in performance-based assessment, including criterion-referenced assessment, competency-based assessment and the like. (McNamara 1996:79)

In addition, the confounding of the method of assessment with the trait being measured is viewed as problematic by many language testing researchers (eg Stevenson 1981; Skehan 1984; Bachman 1990). According to this argument, the indistinguishability of trait and method in the interview means that a rating of speaking ability derived from an interview can be interpreted solely as an

indicator of a person's ability to perform under a particular set of test method conditions (Skehan 1984; Bachman 1990) and is not generalisable beyond the testing situation. In empirical investigations of this issue, Bachman and Palmer (1981, 1982) found that ratings of language abilities based on oral interviews included substantial variance components associated with the test method, in some cases actually larger than those associated with the trait measured.

Uncontrolled factors in the oral interview

Studies of oral interview-based assessments have shown that the sociolinguistic context of the interview, including the skill and experience of the interviewer, may have a significant effect on the outcomes (see, for example, van Lier 1989; Bachman 1990; Ross 1992). Judgments of proficiency may be affected by a wide range of variables such as social or cultural status, age, personality and gender of interlocutors (van Lier 1989, Porter 1991; Wigglesworth 1997a); the topic and purpose of the interaction; the discourse domain (Selinker and Douglas 1985); the medium of exchange (Tarone 1988; Tarone and Yule 1989; Gass et al 1989) and the amount of time available for planning discourse (Wigglesworth 1997b).

All of these uncontrolled factors may affect the amount and quality of the language produced by the candidate during the interview and thus influence the proficiency rating which is given. In 'free-form' interviews such as the ASLPR, the rating awarded may thus depend largely on the rater's behaviour. Whether or not candidates get the chance to produce their 'best' sample of language is thus heavily dependent on the skills of the interviewer (Brown and Lumley 1997, Morton et al 1997).

Collapsing of different aspects of language behaviour

A number of writers have argued that general proficiency rating scales such as the ASLPR and the ACTFL Guidelines fail to reflect the complexity and multidimensionality of language development and use (eg Bachman 1990; North 1995; Brindley 1998b). In such scales, a wide range of language behaviours are collapsed into a single level description which raters are required to match to the testee's performance. However, this creates a problem for raters since particular aspects of a skill may be at different stages of development (eg fluency and pronunciation). This phenomenon is best illustrated in the well-known 'terminal 2+' syndrome (Lowe 1988), a term coined to describe a learner who has a very wide vocabulary but minimal command of syntactic structures.

Ingram (1984a:10) acknowledges this problem when he notes that:

... language is highly complex, that it may develop at slightly different rates in different directions and that, therefore, minor but compensating variations may occur within the total picture of the learner's development.

However, when there are marked discrepancies between an individual's mastery of component skills it becomes very difficult to match performances to a single level description. In this regard, Corbel (1992) broke down the ASLPR descriptors at each level into a set of short statements describing each of the criteria used in the level statement (such as *intelligibility, fluency, vocabulary*). He then asked teachers to rate language samples according to each of these criteria separately rather than using the general level descriptions, using a hypertext-based expert system. Corbel found that an individual ASLPR profile may reflect behaviours from a wide range of different levels (for example, a person who receives an overall rating of 3 may display behaviours corresponding to the descriptions at Levels 2 and 4). This finding calls into question the validity of descriptors which group together different aspects of speaking ability into a single overall level description.

Oversimplification of processes of second language development

Another related issue affecting the construct validity of rating scales such as the ASLPR and ACTFL is the extent to which the proficiency descriptions reflect what is known about the nature of second language development (Brindley 1986, 1991, 1998b; Davies 1992; Lantolf and Frawley 1988; Pienemann et al 1988). Here the developers of the ASLPR make quite strong claims concerning the capacity of rating scales to provide information on universal patterns of language development. For example, Ingram (1995:17) states:

The ASLPR ... seeks to describe the changes that are observable in language behaviour as a learner's proficiency develops from zero to native-like, ie from inability to use the language for any practical purpose to an ability indistinguishable from that of a native speaker of the language.

Ingram (ibid) seems to be suggesting that the ASLPR provides a description of universal second language acquisition processes:

The ASLPR seeks to provide a fairly comprehensive picture of language behaviour related to its view of how interlanguage develops.

However, a number of writers have argued that such claims are overstated, since they are not based on empirical research and do not reflect the current state of knowledge of language development (Davies 1992; McNamara 1996; Brindley 1998b; Pienemann et al 1988). Davies (1992:12–13) comments:

The very strength of the ASLPR, its security through consistency, its safe scaffolding, may persuade us into thinking that proficiency is now all safely tucked up in the ASLPR. That is the danger of over-claiming. Nor does it resolve the doubt about measurement.

Davies (1992:13) invokes Bachman's argument about the context dependence of proficiency descriptions to put forward a case against the over-interpretation of direct tests of language performance such as the ASLPR interview:

What a direct test does is to test specific performance! That is the strong argument against the ASLPR, not against its helpful reminder to us that we should think in explicit terms about proficiency, but against our gradually allowing it to be used as if it were itself a measure, indeed in some contexts the only measure. It isn't and it should not be.

Potential unreliability of single ratings

ASLPR ratings are usually carried out by a single judge. However, there is considerable evidence from both the educational measurement and language testing literature to suggest that single ratings are unreliable even when judges are highly trained (Lunz et al 1990; North 1993; Lumley and McNamara 1995; McNamara 1996). McNamara (1996) reviews a range of studies which have used Rasch measurement to investigate differences in rater severity in assessments of language performance. He reports consistent findings of large differences in severity which do not appear to change in spite of training. This means that a candidate whose performance is judged by a severe as opposed to a lenient rater would be disadvantaged. McNamara concludes (1996:235) that 'assessment procedures in performance tests which rely on single ratings by trained and qualified raters are hard to defend' and advises the use of at least two raters. Similarly, in the context of classroom assessment, Genesee and Upshur (1996:59) state that 'it is best to avoid using students' performance on a single occasion as the sole basis for making decisions about them' and recommend the use of different procedures on different occasions.

Unclear relationship between descriptors and assessment tasks

Alderson (1991:72–74) makes a useful distinction between three functions that proficiency rating scales can serve: the *user-oriented* function where the scales provide information on the meaning of the levels to test users; the *assessor-oriented* function where they are used as rating criteria to assess the quality of a performance and the *constructor-oriented* function where they provide guidance to test constructors.

Pollitt (1991) has argued that one of the problems with scales such as the ASLPR is that they serve the reporting function rather than the assessment function, that is, they do not describe the qualities of a specific performance, but rather provide a general description of what a person at a particular level can do. Thus, at ASLPR 3 Writing, a learner is *able to write with sufficient accuracy in structures and spelling to meet all social needs and basic work needs* (general description). The learner *can write in all those forms used in daily life (personal letters, notes, telegrams, invitations etc) without errors intruding on a native speaker's comprehension and acceptance* (examples of specific tasks). However, these statements give little concrete guidance to the teacher who is developing the assessment tasks which are intended to elicit the behaviour in question. There is very little precise information in either the general level description or the examples of specific tasks which would give assessors an idea of what behaviours to look for or how to assess them against the scale.

Comparability of assessment tasks

As noted above, the ASLPR is not accompanied by a standard set of materials to elicit language samples for assessment. The texts and tasks used by assessors thus vary not only from interviewer to interviewer in the same AMES centre, but also from centre to centre (McIntyre 1993:8). Although examples of tasks are given which might be appropriate for learners at the various levels, there is no indication as to which tasks can be substituted for others. For example, at ASLPR 3 Speaking, a learner *can cope with everyday difficult linguistic situations, such as broken plumbing, a personal misunderstanding, undeserved traffic ticket* etc. However, the sample tasks here could be of a quite different order of difficulty according to the context, the relationships between participants, the goal of the interaction etc. This could unfairly disadvantage learners who are given harder tasks. There is some evidence from a study by Wapshere

(1997) to suggest that this may be the case with the ASLPR. In an item analysis of tests based on ASLPR reading tasks, she found that six out of eight pairs of reading tasks considered to be appropriate for Level 2 were of unequal difficulty — in some cases dramatically so — even though the paired tasks were based on the same text type and judged by experienced teachers to be of equivalent complexity.

In the absence of any guidance on either the specific characteristics of particular texts or on the types of test items which could be used to elicit these behaviours, there is scope for a good deal of variation in the types of assessment tasks and materials which teachers use. On this point, Quinn and McNamara (1987:8) comment:

... as long as the individual rater is free to substitute any task for the example tasks given, the whole process comes close to being the use of a variable instrument, a possible consequence of which is variable measurements.

Task quality

It follows from the previous point that the validity and reliability of the ASLPR depend to a large degree on the quality of the assessment tasks that teachers create. In the case of reading and listening assessment, this may place considerable demands on teachers' skills in assessment task design since not only do they have to select appropriate texts but they also have to devise questions or items which will elicit the behaviours described in the scale. This is akin to developing a 'mini-test' consisting of one or more passages along with rubrics and sets of items, an activity which requires a good deal of time and skill if the information provided by the test is to be dependable.

The ASLPR: Summary

Although it has been in use for nearly twenty years within the AMEP and appears to serve some of its purposes quite adequately (McIntyre [1995], for example, reports an 85% success rate in placing candidates in classes of the appropriate level), a number of questions continue to surround the ASLPR. These concern the validity of the scale descriptors, the adequacy of the oral interview procedure as a measure of spoken language ability and the comparability of the teacher-developed assessment tasks which are used to place learners on the scale.

The Certificates in Spoken and Written English (CSWE)

Background

Competency-based approaches to language education have their origins in the large-scale attempts which have been made over recent years in a number of industrialised countries such as the United Kingdom, Australia, Canada and New Zealand to reform the education and training systems. The cornerstone of these reforms is the belief that a clear articulation of the skills and knowledge involved in occupational performance and the setting of competency standards will facilitate curriculum development, improve the quality of learning and ultimately lead to an increase in productivity and international competitiveness.

As it is currently used in the education and training literature in the UK and Australia, the term *competency* refers to 'the ability to perform the activities within an occupation or function to the standard expected in employment' (National Training Board 1991:30). Competency in this sense may include any attributes that contribute to work performance. Competency statements involve the specification of the knowledge and skill necessary in occupational performance and the application of that knowledge or skill.

Competency-based models of vocational education and training have in recent years begun to dominate the educational landscape in Australia, the UK and New Zealand. They are also beginning to become firmly established in the field of adult language learning as a basis for curriculum design, assessment and reporting. In the UK, the Royal Society of Arts (1987, 1988) developed profile schemes in practical skills and ESL which were based on explicit specifications of learning outcomes in the form of Profile Sentences (stated in terms very similar to what are now called competencies or units of learning) and which incorporated various forms of criterion-referenced assessment. More recently, also in the UK, the Languages Lead Body (LLB) (1993) has produced a competency-based National Language Standards framework for the use of foreign languages at work, aimed at providing a 'common specification which can be used by training providers, employers and individuals when considering levels of language skills and training required'.

Competency-based principles also underpin the CSWE, the curriculum framework used within the AMEP and described briefly earlier in this chapter. The framework describes learning outcomes at four levels in terms of *language competencies*, typically expressed as 'can-do' statements (*can participate in a casual conversation; can respond to spoken instructions; can read an informa-*

tion text; can write a report etc). Each competency is broken down into *elements* which describe the skills and knowledge involved in the performance, the *performance criteria* which set out the obligatory elements of performance against which the learner is to be assessed and *range statements*, which specify the conditions under which the performance occurs (eg the time available for the task, the amount of support and resources available etc). Examples of sample tasks which might be used to gather evidence of successful performance are also included. An example of a CSWE competency is given in Figure 2.

Certification and assessment in the CSWE

The award of a Certificate is contingent on the learners' attainment of a given number of competencies in each language skill, the total number of which may range from thirteen to sixteen, depending on the level. Assessment is usually conducted by class teachers using their own assessment tasks or sample tasks which accompany the Certificate (NSW AMES 1998). A set of assessment support materials has also been developed (Christie and Delaruelle 1997).

The rating scheme used requires teachers to make a binary decision on whether or not each of the mandatory performance criteria in a given competency has been achieved. In order to qualify for the award of a competency, all of the performance criteria must be fulfilled. Teachers are advised to assess students when they believe the students are familiar with the competency to be assessed. Students who fail to achieve a competency are given the opportunity to re-attempt the same competency later. (NSW AMES 1998:37)

Theoretical basis

The theoretical rationale of the CSWE derives from Halliday's (1985) systemic-functional linguistic theory 'which systematically relates language to the contexts in which it is used' (Hagan et al 1992:1) and draws on applications of genre theory in other educational contexts (see Martin 1993 for an overview). The basis of the competency specifications is the language user's knowledge of the relationship between text and context which is used 'to predict the language likely to be used in any given situation' (NSW AMES 1998:18). The text is seen as an encounter involving a number of successive stages in which certain linguistic features are mobilised to achieve the social purpose of the task. The competency descriptions describe the relationships between text and context in terms of the three Hallidayan components of register-*field* (the nature of the action that is taking place), *tenor* (participants and their relationships) and

Elements	Performance Criteria	Range Statements	Evidence Guide
<ul style="list-style-type: none"> i. can use appropriate strategies to negotiate transaction ii. can use appropriate vocabulary iii. can request information iv. can provide information 	<ul style="list-style-type: none"> • uses appropriate strategies to negotiate transaction eg opening and closing, confirming, checking • uses appropriate vocabulary • requests information as required using questions or statements • provides relevant information 	<ul style="list-style-type: none"> • at least 1 minute in length • familiar and relevant • 2 speakers • sympathetic interlocutor • face-to-face/telephone • telephone for Distance Learning students • may include a few grammatical or pronunciation errors but errors should not interfere with meaning or dominate text • recourse to clarification/repetition 	<p>Sample tasks</p> <ul style="list-style-type: none"> • Learners ask about enrolling in a class • Learners ask about child-care or enrolling child in school

Figure 2: Example of CSWE competency — Can negotiate an oral transaction to obtain information (Certificate II, Competency 6)

Phonology

It is assumed that:

- articulation of some phonemes and clusters as well as intonation, stress and rhythm in longer phrases and clauses may often be inaccurate or unconventional
- teaching programs will pay attention to:
 - phonological features of longer utterances
 - developing learner self-monitoring and repair/correction strategies

Source: NSW AMES 1998

mode (the role played by language). The elements and performance criteria are described in terms of *discourse structure* (eg the sequencing of information, cohesion, reference) and *grammar and vocabulary* (the features relevant to the text concerned).

The CSWE: Issues and problems

Validity

According to Messick (1989), validity considerations encompass the extent to which the construct that is defined reflects current theoretical understandings of the attributes or qualities being tested and, by extension, the extent to which the tasks adequately sample the domain of interest (language proficiency). In addition, the social consequences in terms of consequences and fairness need to be taken into account. As yet, there has been relatively little research into any of these aspects of the CSWE. A number of general observations can nevertheless be made on the basis of available evidence.

First, it is important to note that validity judgments will depend on the way that those designing the assessment procedures choose to define the construct in question — in this case language ability. The construct definition will determine which competencies are identified, the terms in which they are described and, by extension, the nature of the tasks and of the criteria which are used to assess performance. Consequently, differing construct definitions are likely to yield quite different assessment criteria. In relation to the CSWE, Brindley (1994) shows how the elements defined in the CSWE III oral competency, *can negotiate complex/problematic exchanges*, differ dramatically from those described in a similar competency (*can participate effectively in negotiation*) in the Royal Society of Arts Practical Skills Profile Scheme (RSA 1988) which provides a method for assessing work-related and non-vocational courses in Communication, Numeracy and Process Skills. He concludes:

The lack of commonality between these two sets of criteria illustrates graphically how the perspective of the test designer can influence the way that competencies are described and hence assessed. In the first case, it is primarily language that is the object of assessment (staging of discourse, conversational strategies, information-giving etc). In the second, it is interpersonal communication skills in a more general sense and non-linguistic, social and affective factors (intentionality, sympathy, empathy) have a much greater role. (Brindley 1994:45)

Brindley (op cit p 46) argues that if the CSWE is to be used to certify competency in 'real life' communication outside the classroom, then there may be a case for adopting a broader definition of language ability than has conventionally been the case in language assessment. This is consistent with the call by Shohamy (1995) and McNamara (1996) for an expanded model of language performance which can encompass non-linguistic factors.

Another key issue affecting the construct validity of the CSWE is the relationship between construct definition and scoring methods. In competency-based assessment, the assessor is required to make a binary judgment as to whether or not the learner demonstrates the behaviours specified in the performance criteria. Evidence that the learner has fulfilled each of the performance criteria must be present before the learner can qualify for the award of the competency. In the case of the CSWE, the performance criteria according to which the quality of a language performance is evaluated consist of certain obligatory features of oral and written discourse.

However, there is considerable doubt surrounding the extent to which it is possible (or desirable) to specify the mandatory elements of communicative acts:

It may well be that some aspects of human behaviour, skills and knowledge are so conventionalized that they lend themselves to competency descriptions and somewhat ritualized performance testing. But there is a strong case for saying that human language, because of its inherent creativity, cannot be so characterized without serious, reductionist distortion. Complex skills like language must be more than an amalgamation of purportedly separable parts. (Quinn 1993:72)

In this connection, Bottomley et al (1994:21) comment that the 'presentation of text types (in the CSWE) tends to gloss over the difficulty of capturing with certainty the elements of (particularly spoken) discourse', while Kress (1993:28) argues that genres are not fixed and in constant evolution:

... while generic conventions provide certain dimensions of constraint, generic form is never totally fixed, but is always in the process of change — for example, a job interview in 1992 is very different from a job interview in 1932.

Murray (1994:63) comments that the text-based approach relies on the availability of very comprehensive descriptions of different oral and written genres but notes that 'full descriptions of the structures of most oral and written gen-

res have yet to be developed'. For this reason she uses criteria for portfolio assessment of student writing in the form of an assessment guide based on the more 'subjective, general features of written texts' (ibid) rather than on genre-specific features.

Quinn (1993:81) argues that evidence for the construct validity of the competencies is lacking and calls for research which would establish:

... what the evidence is that particular tasks exemplify or instantiate or demonstrate particular elements of competencies, and what the evidence is that the stated elements actually define the competence.

In order to validate the competency statements a good deal of further research needs to be undertaken into the relationship between actual language performance and the features that make up the elements and performance criteria. Such research would help to establish to what extent certain performance features are obligatory as opposed to optional and thus help to inform the criteria that are used to evaluate communicative performances in the CSWE.

Reliability

A number of studies have investigated the issue of rater consistency in the CSWE. These include a study by Jones (1993) who investigated competency ratings for Stage 2 (now Certificate II) of the first version of the CSWE. She found quite high levels of overall rater agreement (calculated in terms of percentages) for most competencies. The study also yielded information on the way in which teachers were interpreting and applying the performance criteria which was subsequently used to modify the criteria in later versions of the Certificate. For example, the competency *can understand and give spoken instructions* proved problematic to assess for some teachers since it requires the simultaneous assessment of listening and speaking. At the same time, assessors reported that they found some of the relativistic terminology used to describe performance ('mostly appropriate', 'mostly correct spelling' etc) difficult to interpret. Another interesting outcome of this investigation related to the relationship between the nature of the assessment task and the type of response produced by learners. Teachers reported that some texts were difficult to assess because they did not conform to the specified text type ('this task does not generate a recount') and thus could not be marked according to the given performance criteria.

Aldred and Claire (1994) investigated the reliability of rater classifications of

competency achievement in thirteen reading and writing competencies and five oral competencies from Stage 3 of the first version of the CSWE. Using Cohen's *kappa*, an index of classification consistency that compensates for chance agreement, they found widely varying levels of agreement on ratings of the performance criteria. However, rater agreement improved when benchmark performance samples and detailed indicators for particular performance criteria were supplied. In their list of recommendations they suggested that:

- performance criteria which are identified as problematic should be omitted or revised
- a system of weighting criteria or of making some criteria optional should be investigated
- criteria or range statements which use generalised descriptions such as *mostly correct spelling ...* or *grammatical errors do not interfere with meaning* should be illustrated with reference to benchmark performances or specify actual error rates
- descriptions of performance criteria which occur in several competencies should be standardised
- detailed answer keys should be developed for comprehension tasks.

In addition they recommended the development of a set of task design guidelines to support teachers in designing additional tasks, along with an assessment task bank which could be drawn on by teachers. The importance of ongoing empirical investigation of assessment practices was also highlighted.

A number of these recommendations were taken into account in the preparation of the 1995 edition of the Certificate (NSW AMES 1995). Some competencies were rewritten and performance criteria were modified or deleted in the light of feedback from raters. A set of guidelines to assist teachers in task design was also produced (Christie and Delaruelle 1997). Further modifications were subsequently made and an updated edition issued in 1998.

Transfer and generalisability

The issue of generalisability is not a straightforward one as far as the CSWE is concerned. Although the preface to the first edition of the CSWE is not explicit on this point, it states that the competencies 'have been designed to operate at the level of the curriculum and are therefore expressed in general terms' (Hagan et al 1992:33), which would appear to imply some kind of transferabil-

ity. The way in which the competencies are described ('can write short reports' etc) would, similarly, give the impression that a single performance could be regarded as an indicator of a more general ability which would enable the learner to perform a similar task on another assessment occasion.

In the second edition of the CSWE (NSW AMES 1995) however, competencies are expressed in terms of the learner's performance on a single task on a single occasion ('can write a short report' etc). In addition, the guidelines for task design which accompany the CSWE (Christie and Delaruelle 1997) state:

Assessment within a competency-based curriculum framework is known as achievement assessment. This means that the assessment is concerned with what a student can do in relation to a given syllabus or course of study. In other words, the assessment is concerned with whether a student has learned X or Y in a given course of study and can perform to a specified standard. This type of assessment is not oriented to future abilities. It does not, for example, predict whether a student will pass a university entry test or be successful in a future course of study or employment. This is not to say that inferences cannot be drawn from the outcomes of achievement assessment, but rather, that specific claims are not being made. (p 2, emphasis added)

Elsewhere Burrows (1995) writes:

Achievement assessment measures student use of familiar learnt language in familiar, learnt and relevant situations, within the classroom or outside it ... Students learn to use language in the classroom and assessment measures what they have learnt. It does not measure the proficiency they have in the language. (p 33, emphasis added)

Statements such as this suggest that the CSWE assessments are not intended to be generalisable beyond the assessment situation. At the same time, however, the door seems to be left open to the possibility that transfer may take place.

Grove (1996:13) points out that this narrow conceptualisation of achievement may cause problems with external audiences who want information on proficiency outcomes:

Specified achievement, not proficiency, is CSWE's goal; a learner is accredited with a competency once he or she has demonstrated success on a single task on a single occasion. There is no further investi-

gation. And here we come to a major paradox of the situation, for the Government wants language development to be one of the ongoing and transferable skills of the workplace, yet the competency-based assessment procedures of CSWE confirm only that a learner has fulfilled a task; they have nothing to say about the learner's capacity to transfer those skills to another context. Indeed, the authors of the CSWE curriculum and assessment documents see this as a positive feature of the Certificate: it validates itself by claiming to measure no more than achievement in the course of study it has set.

It should be noted here that the rejection of task generalisability is, on the face of it, consistent with Halliday's explicit rejection of any distinction between underlying knowledge and realisation of that knowledge. Halliday argues (1970:145) that:

... linguistics is concerned with the description of speech acts, or texts, since only through the study of language in use are all the functions of language, and therefore all components of meaning, brought into focus. Here we shall not need to draw a distinction between an idealized knowledge of a language and its actualized use: between the 'code' and 'the use of the code', or between 'competence' and 'performance'. Such a dichotomy runs the risk of being either unnecessary or misleading: unnecessary if it is just another name for the distinction between what we have been able to describe in the grammar and what we have not, and misleading in any other interpretation. The study of language in relation to the situations in which it is used — to situation types, i.e. the study of language as 'text' — is a theoretical pursuit, no less interesting and central to linguistics than psycholinguistic investigations relating the structure of language to the structure of the human brain.

On the other hand, as McNamara (1996:88) points out, a distinction between meaning potential (the semantic system) and actual realisation in instances of use ('text') is central to systemic theory. In this sense, the CSWE competencies may reflect a notion of 'generic competence', although this is not stated explicitly. This is what could be understood from a reading of the guidelines concerning reassessment of competencies which a student may previously have failed to achieve:

Reassess the students who did not achieve the competency later in the

course. Do not use the same task but assess the skills in a different context. (NSW AMES 1998:37)

Elsewhere, describing different approaches to assisting English language learners in the workplace to gain control of different registers in English, Joyce (1992:26, emphasis added) writes in relation to the text-based approach which informs the CSWE:

In teaching this type of text, teachers have a number of choices. These choices include explaining the surface features of vocabulary or rewriting the text. A third choice is to use the text to develop text knowledge and transferable skills which students can use when they engage with similar texts in the future. This approach involves modelling for the students how information in this instructional text is structured into the text of technical texts in the workplace. (p 26, emphasis added)

Such statements would seem to imply some sort of skills transfer within the same competency. However, there has been no research into the extent to which the skills deployed by a learner in achieving a CSWE competency transfer either to other competencies or to other tasks assessing the same competency. In the meantime, in the absence of such information, it must be assumed that the inferences which can be made on the basis of CSWE performances are very restricted, as a number of commentators have pointed out (Quinn 1993; Cope et al 1994; Grove 1996).

Task comparability

Another key aspect of generalisability is the replicability of tasks. Since Certificate attainment is determined on the basis of teacher-developed assessment tasks, the quality of these tasks will have a major effect on the dependability of the information on learner outcomes. Scores must have the same meaning across different tasks and contexts; consequently, task-related and rater-related variability should desirably be kept to a minimum. This means that tasks which are aimed at assessing the same competency should be parallel, that is, they should elicit the same range of language features under the same conditions of administration, they should present a comparable challenge for learners and they should be rated in the same way.

The issues of task comparability in the CSWE are similar to those which were raised in relation to the ASLPR. However, in the case of the CSWE, the chal-

length of developing tasks is even greater since the teacher must make sure that the tasks elicit those specific features of language which are described in the performance criteria. If these features are not present, then the student will not be able to achieve the competency. This places considerable demands on teachers' skills in assessment task design.

The CSWE: Summary

The CSWE offers a way of closely integrating instruction, assessment and reporting. The use of competencies as the unit of instruction and assessment allows gains to be reported which might not be detectable using a general proficiency scale, thus giving a more accurate picture of individual achievement. In addition, the specification of explicit performance criteria enables teachers to give diagnostic feedback to learners following assessment. However, as with the ASLPR, there is little evidence to support the validity and reliability of the CSWE assessments. The question of the generalisability of competency achievement also remains open to question.

AMEP assessment: Research issues

The previous discussion has identified a number of important issues and questions concerning the nature of language performance assessment and the way in which it is conducted in the AMEP. These issues provided the stimulus for the research studies reported in this volume and are summarised briefly below. Given the central role of the CSWE as the mandated national framework for curriculum and assessment, four of the five studies reported here are concerned with aspects of competency-based assessment involving the qualities of the CSWE assessment tasks and the way in which they are used. For outcomes of a range of research projects into the *access:* test, interested readers should consult Brindley and Wigglesworth (1997). Information and research relating to the ASLPR is reported in Ingram (1990, 1984a and b, 1995).

1 The relationship between the CSWE and ASLPR

CSWE Certificate levels and ASLPR levels are used interchangeably for placement and reporting purposes. As noted above, inferences concerning the relationship between competency attainment and proficiency levels are made on the basis of these assumed equivalences. However, the relationship between the CSWE and ASLPR has never been empirically investigated. Although the scores on one assessment cannot be converted directly into the other, an examination

of the comparability of the tasks used and the skills assessed would help to ascertain the extent to which the comparisons are valid.

This issue is addressed in Chapter 2 by Geoff Brindley who reports on the results of a detailed examination of the content of the reading assessment tasks and items used in *access:*, ASLPR and CSWE. On the basis of a comparative study of the content of the different assessments, the author finds that the tests differ along a number of dimensions and concludes that *access:* is the most difficult test of the three. In general, the CSWE reading tasks appear to make greater demands on learners than the ASLPR tasks, suggesting that the current practice of equating ASLPR Level 2 with CSWE Level III may not have any basis in fact. However, Brindley suggests that this issue needs to be investigated further through a comparative examination of performances of a common cohort of learners in all four skills. He also identifies a number of methodological problems surrounding the use of content rating instruments which rely on expert rater judgments and concludes that expert judgments need to be supplemented by reports of the actual test-takers themselves.

2 Task comparability and task difficulty

ASLPR and CSWE assessments both elicit performances which are rated against a set of external standards in the form of scale descriptors or competency statements. However, in the absence of standardised assessment tasks, it is up to individual teachers to design their own assessments, leaving scope for considerable variability in task design and administration. Here the key issues are whether assessment tasks aimed at assessing the same skill (in the case of the ASLPR) or the same competency (in the case of the CSWE) elicit similar performances and are of comparable difficulty. The answers to these questions are crucial in ensuring the validity and dependability of the information yielded by the assessment tasks.

Chapter 3 and Chapter 4 both deal with these questions. In Chapter 3 Gillian Wigglesworth examines the way in which variations in task conditions and task characteristics influence the assessment of spoken language competencies in the CSWE and explores the implications for assessment task design practice. She finds considerable variability in both learner and interlocutor behaviour within the same assessment tasks and concludes that even relatively small changes in the characteristics and/or conditions of the Task may significantly affect task difficulty. These findings suggest that it is important to draw up precise task specifications at the design stage and to trial oral tasks with a range of

different learners and interlocutors before using them for assessment purposes. Wigglesworth also considers the implications of her findings for teacher professional development and emphasises the need to raise teachers' awareness of the influence that they may have on learners' output in oral assessment tasks. She concludes with a call for ongoing research into factors affecting assessment task performance.

In Chapter 4 Geoff Brindley reports on an investigation into task difficulty and task generalisability in six CSWE writing competency assessment tasks. Using many-faceted Rasch analysis and generalisability theory, he finds that assessment tasks aimed at tapping the same competencies are of different levels of difficulty but that these differences are not substantial. On the other hand, raters appear to differ markedly in severity. The analysis also reveals very low reliability for single ratings. On the basis of this finding, the author suggests that increasing the number of raters would improve decision dependability. Investigating the question of generalisability of skills across tasks assessing the same competency, he finds that 'generic' writing ability components do not appear to transfer across different writing tasks. He hypothesises, however, that these results are more likely to be due to a 'halo effect' in the rating process rather than an indicator of the task-specificity of writing skills. In a discussion of the implications of the study for assessment policy and practice he recommends, like Wigglesworth in Chapter 3, that tasks need to be very carefully designed and trialled. In conclusion, Brindley highlights the need for educational institutions to provide adequate professional development support for teacher assessment.

3 Rater consistency

All of the assessment procedures described in this chapter rely on subjective judgments of language performance. However, as noted in the previous discussion, these judgments may be quite variable across different raters and different occasions. In order to ensure fairness to students, it is important to constantly monitor the consistency with which ratings are administered. This involves not only ascertaining the extent to which raters agree on the overall quality of learner performance samples (see Brindley, Chapter 4) but also the extent to which they interpret the assessment criteria in the same way.

Chapter 5 by David Smith explores the question of rater judgments in the CSWE. Using think-aloud protocols, he analyses the way in which raters interpret and apply the CSWE performance criteria when assessing writing texts.

He finds generally high levels of agreement amongst raters on whether or not criteria for competency achievement have been met but they appear to interpret the individual performance criteria in different ways. However, despite this variation, the raters seem to rely on the given performance criteria as a basis for making their decisions. Smith's analysis of the verbal protocols identifies three different reading strategies used by raters to assess a text and reveals a close relationship between rater judgments and the strategies employed. On the basis of the results of the study, the author suggests ways in which the CSWE competency specifications could be modified to reduce ambiguity of interpretation. He concludes with an examination of implications for professional development and further research.

4 Reporting aggregate competency outcomes

In a climate in which all government programs in Australia are required to demonstrate that they are achieving targeted outcomes, the question of how to report aggregate competency gains across the AMEP is becoming increasingly important. DIMA has accordingly commissioned a number of studies into AMEP outcomes with a view to developing 'benchmarks' that reflect the gains which can realistically be expected within a given period of tuition (Ross 1997, 1998). However, estimates of likely gains derived from aggregate data are likely to be misleading if they do not take into account the differences which exist between AMEP clients. In a study of the competencies achieved by over 10 000 AMEP learners in 1995–96, Ross (1997) found that learners who enter the AMEP at the same proficiency level may achieve quite different outcomes. This suggests that factors other than proficiency have a major impact on competency gains and highlights the need to identify those characteristics which most inhibit or promote learning. Information of this kind can facilitate the placement of learners into learning programs that match their individual characteristics. At the same time, it can also assist program managers and policy makers to estimate the amount of learning time needed by learners with particular profiles.

This issue of expected gains is taken up in Chapter 6 by Steven Ross who examines the influence of a number of individual difference variables on CSWE competency achievement. The variables under investigation include age, sex, educational level, first language and length of residence in Australia. Using a range of quantitative modelling techniques, Ross demonstrates that education in home country is the most consistent predictor of competency outcomes, followed by age on arrival. He suggests that information on these factors along

with language proficiency level would be minimally sufficient to optimally place clients. Ross concludes by canvassing the possibility of a computerised procedure using an expert system which could be used to classify AMEP learners more accurately at entry to the Program.

Conclusion

The research reported in this volume range spans a wide range of assessment concerns, ranging from issues of macro-level reporting through to the functioning of individual performance criteria. It is hoped that the chapters in combination succeed in giving the reader a sense of the complexities involved in implementing and monitoring an assessment and reporting system that can meet the needs of decision makers, teachers and learners themselves while at the same time meeting technical requirements of validity and reliability.

The findings of the various studies raise a number of issues of significance in relation to the development and use of performance assessments in language learning programs and suggest a range of research directions which could be further pursued within the program under investigation, the AMEP. These could include:

- validation studies of the CSWE competencies which would look at the relationship between the performance criteria and the performances produced in response to CSWE assessment tasks (Fulcher 1996b);
- further studies of the way in which raters interpret and apply assessment criteria using both introspective and retrospective techniques (Vaughan 1991; Delaruelle 1997);
- investigations of the types of variables which affect task difficulty and their relative effects on task performance, both singly and in combination (Fulcher 1996a; Brindley 1998a);
- studies of the demands which assessment task construction and administration make on teachers, in terms of time, skills and resources; and
- studies of the ways in which information on learner outcomes is interpreted and used by policy makers.

In conclusion, it is worth highlighting a key element in the implementation of performance assessments such as those described in this volume, and that is teacher professional development. It is apparent from many of the studies reported here that the success of outcomes-based assessment systems such as

the CSWE framework will depend heavily on the qualities of the assessment tasks developed by teachers. As Brindley points out (Chapter 4, this volume), a continuing commitment to professional development on the part of both educational authorities and teachers themselves will be required to ensure that the assessment system provides high quality information to all stakeholders.

Notes

- 1 The level descriptors used in the ASLPR have undergone revision in recent years. The scale is now known as the ISLPR (International Language Proficiency Ratings) (Ingram and Wylie 1997). However, it is the original 1984 version which is still used in the AMEP.
- 2 An overall band score of 5 with a minimum score of 5 in each component of IELTS is now required to meet the vocational proficiency requirement.

References

- Alderson, J C 1991. Bands and scores. In J C Alderson and B North (eds). *Language testing in the 1990s*. London: Macmillan, 71–86
- Aldred, M and S Claire 1994. *Investigating consistency in assessment at Stage 3 of the Certificate in Spoken and Written English*. Sydney: NSW Adult Migrant English Service
- Bachman, L F 1990. *Fundamental considerations in language testing*. Oxford: Oxford University Press
- Bachman, L F and A S Palmer 1981. A multi-trait multi-method investigation into the construct validity of six tests of speaking and reading. In A S Palmer, P J M Groot and G A Trosper (eds). *The construct validation of tests of communicative competence*. Washington, DC: TESOL, 149–65
- Bachman, L F and A S Palmer 1982. 'The construct validation of some components of communicative proficiency'. *TESOL Quarterly* 16, 4: 449–65
- Bottomley, Y, J Dalton and C Corbel 1994. *From proficiency to competencies: A collaborative approach to curriculum innovation*. Sydney: National Centre for English Language Teaching and Research, Macquarie University
- Brindley, G 1986. *The assessment of second language proficiency: Issues and approaches*. Adelaide: National Curriculum Resource Centre

- Brindley, G 1991. Defining language ability: The criteria for criteria. In S Anivan (ed). *Current developments in language testing*. Singapore: Regional Language Centre, 139–64
- Brindley, G 1994. 'Competency-based assessment in second language programs: Some issues and questions'. *Prospect*, 9, 2: 41–55
- Brindley, G 1997a. 'Assessment and the language teacher: Trends and transitions'. *The Language Teacher*, 21, 9: 37, 39
- Brindley, G 1997b. Investigating second language listening ability: Listening skills and item difficulty. In G Brindley and G Wigglesworth (eds). *access: Issues in language test design and delivery*. Sydney: National Centre for English Language Teaching and Research, Macquarie University, 65–86
- Brindley, G 1998a. 'Outcomes-based assessment and reporting in second language learning programs: A review of the issues'. *Language Testing*, 15, 1: 45–85
- Brindley, G 1998b. Describing language development? Rating scales and second language acquisition. In L F Bachman and A D Cohen (eds). *Interfaces between second language acquisition and language testing research*. Cambridge: Cambridge University Press
- Brindley, G and G Wigglesworth (eds) 1997. *access: Issues in language test design and delivery*. Sydney: National Centre for English Language Teaching and Research, Macquarie University
- Brown, A and T Lumley 1997. Interviewer variability in specific-purpose language performance tests. In A Huhta, V Kohonen, L Kurki-Suonio and S Luoma (eds). *Current developments and alternatives in language assessment*. Jyväskylä: University of Jyväskylä, Centre for Applied Language Studies, 137–150
- Burrows, C 1994. 'The AMEP meets CBT: A literature review'. *Prospect*, 9, 2: 18–29
- Burrows, C 1995. 'Why assessing oral language is not a waste of time'. *Interchange*, 23: 32–4
- Christie, J and S Delaruelle 1997. *Assessment and moderation: Book 1. Task Design*. Sydney: National Centre for English Language Teaching and Research, Macquarie University

- Cope, B, N Solomon, C Kapitzke, D Plimer, T McNamara, A Brown, H Scheeres, D Slade and K Hill 1994. *Assessment and moderation processes in adult literacy and adult ESL in tendered labour market programs* (1994 Draft Report). Canberra: Commonwealth Department of Employment, Education and Training
- Corbel, C 1992. 'Improving global rating reliability with a category-oriented, Hypertext-based computerised profiling instrument'. Unpublished MA thesis. University of Melbourne
- Davies, A 1992. 'Is language proficiency always achievement?' *Melbourne Papers in Language Testing*, 1, 1: 1–11
- Delaruelle, S 1997. Text type and rater decision-making in the writing module. In G Brindley and G Wigglesworth (eds). *access: Issues in language test design and delivery*. Sydney: National Centre for English Language Teaching and Research, Macquarie University, 215–42
- Department of Immigration and Ethnic Affairs (DIEA) 1995. *Adult Migrant English Program Handbook*. Canberra: Department of Immigration and Ethnic Affairs
- Fulcher, G 1996a. 'Testing tasks: issues in task design and the group oral'. *Language Testing*, 13, 1: 23–52
- Fulcher, G 1996b. 'Does thick description lead to smart tests? A data-based approach to rating scale construction'. *Language Testing*, 13, 2: 208–38
- Gass, S, C Madden, D Preston and L Selinker (eds) 1989. *Variation in second language acquisition: Discourse and pragmatics*. Clevedon, Avon: Multilingual Matters
- Genesee, F and J Upshur 1996. *Classroom-based evaluation in second language education*. Cambridge: Cambridge University Press
- Grove, E 1996. 'Creativity and accountability in competency-based language programmes: Issues of curriculum and assessment'. *Melbourne Papers in Language Testing*, 5, 1: 1–18
- Hagan, P, S Hood, E Jackson, M Jones, H Joyce and M Manidis 1992. *Certificate in Spoken and Written English*. Second Edition. Sydney: NSW Adult Migrant English Service and the National Centre for English Language Teaching and Research

- Halliday, M A K 1970. Language structure and language function. In J Lyons (ed). *New horizons in linguistics*. Harmondsworth: Penguin
- Halliday, M A K 1985. *An introduction to functional grammar*. London: Edward Arnold
- Harris, D P and L A Palmer 1986. *Comprehensive English language test*. New York: McGraw-Hill
- Hawthorne, L 1996. 'The politicisation of English: The case of the *step* test and the Chinese students'. *Australian Review of Applied Linguistics*, Series S, 13: 13–32
- Hill, K 1997. The role of questionnaire feedback in the validation of the oral interaction module. In G Brindley and G Wigglesworth (eds). *access: Issues in language test design and delivery*. Sydney: National Centre for English Language Teaching and Research, Macquarie University, 147–74
- Ingram, D 1981. The Australian Second Language Proficiency Ratings. In J S Read (ed). *Directions in language testing*. Singapore: RELC, 108–38
- Ingram, D E 1984a. Introduction to the ASLPR. In *Australian Second Language Proficiency Ratings*. Canberra: Australian Government Publishing Service
- Ingram, D E 1984b. *Report on the Formal Trialling of the Australian Second Language Proficiency Ratings (ASLPR)*. Canberra: Department of Immigration and Ethnic Affairs
- Ingram, D E 1990. The Australian Second Language Proficiency Ratings (ASLPR). In J H A L de Jong (ed). *Standardization in language testing*. Amsterdam: Free University Press, 46–61
- Ingram, D E 1995. 'Scales'. *Melbourne Papers in Language Testing*, 4, 2: 12–29
- Ingram, D and E Wylie 1984. *Australian Second Language Proficiency Ratings*. Canberra: Australian Government Publishing Service
- Ingram, D and E Wylie 1997. *International Second Language Proficiency Ratings*. Nathan, Queensland: Griffith University
- Jones, M 1993. 'Investigating consistency in assessment in the Certificate in Spoken and Written English'. Draft project report. Sydney: NSW Adult Migrant English Service

- Joyce, H 1992. 'Literacy pedagogy in the context of industrial change — dependency or choice?' *Critical Forum*, 1, 2: 18–29
- Kellett, M R and J J Cumming 1995. 'The influence of English language proficiency on the success of non-English speaking background students in a TAFE vocational course'. *Australian and New Zealand Journal of Vocational Education Research* 3,1: 69–86
- Kress, G 1993. Genre as social process. In B Cope and M Kalantzis (eds). *The power of literacy: A genre approach to teaching writing*. London: The Falmer Press, 22–37
- Languages Lead Body 1993. *Introduction to the national language standards*. London: Languages Lead Body
- Lantolf, J and W Frawley 1988. 'Proficiency: Understanding the construct'. *Studies in Second Language Acquisition*, 10, 2: 181–95
- Lee, T 1995. 'A many-faceted Rasch analysis of ASLPR ratings'. Unpublished manuscript. Nathan: Centre for Applied Linguistics and Languages, Griffith University
- Linn, R L 1994. 'Linking results of distinct assessments'. *Applied Measurement in Education*, 6, 1: 83–102
- Lowe, P 1988. The unassimilated history. In P Lowe and C W Stansfield (eds). *Second language proficiency assessment: Current issues*. Englewood Cliffs, New Jersey: Prentice Hall Regents, 11–52
- Lumley, T J N and T F McNamara 1995. 'Rater characteristics and rater bias: Implications for training'. *Language Testing*, 12, 1: 54–71
- Lunz, M E, B D Wright and J M Linacre 1990. 'Measuring the impact of judge severity on examination scores'. *Applied Measurement in Education*, 3, 4: 331–45
- Manidis, M and P Prescott 1994. *Assessing oral language proficiency*. Sydney: National Centre for English Language Teaching and Research, Macquarie University
- Martin, J R 1993. 'Genre and literacy — modelling context in educational linguistics'. *Annual Review of Applied Linguistics*, 13: 141–72

- McIntyre, P 1993. 'The importance and effectiveness of moderation training on the reliability of teacher assessments of ESL writing samples'. Unpublished MA thesis, University of Melbourne
- McIntyre, P 1995. Language assessment and real-life: The ASLPR revisited. In G Brindley (ed). *Language assessment in action*. Sydney: National Centre for English Language Teaching and Research, Macquarie University, 113–44
- McNamara, T F 1996. *Second language performance testing: Theory and research*. London: Longman
- McNamara, T F and B K Lynch 1997. A generalizability theory study of ratings and test design in the writing and speaking modules of the *access* test. In G Brindley and G Wigglesworth (eds). *access: Issues in language test design and delivery*. Sydney: National Centre for English Language Teaching and Research, Macquarie University, 197–214
- Messick, S J 1989. Validity. In R L Linn (ed). *Educational measurement*. New York: Macmillan, 13–103
- Morton, J, G Wigglesworth and D Williams 1997. Approaches to the evaluation of interviewer behaviour in oral tests. In G Brindley and G Wigglesworth (eds). *access: Issues in language test design and delivery*. Sydney: National Centre for English Language Teaching and Research, Macquarie University, 175–96
- Murray, D 1994. 'Using portfolios to assess writing'. *Prospect*, 9, 2: 56–69
- National Training Board 1991. *National Competency Standards*. Canberra: National Training Board Ltd
- New South Wales Adult Migrant English Service (NSW AMES) 1995. *Certificates in Spoken and Written English*. 2nd ed. Sydney: New South Wales Adult Migrant English Service
- New South Wales Adult Migrant English Service 1998. *Certificates in Spoken and Written English, I, II, III and IV*. Updated edition. Sydney: New South Wales Adult Migrant English Service
- North, B 1993. *The development of descriptors on scales of language proficiency*. Washington, DC: The National Foreign Language Center

- North, B 1995. 'Scales of language proficiency'. *Melbourne Papers in Language Testing*, 4, 2: 60–111
- O'Loughlin, K 1995. 'Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test'. *Language Testing*, 12, 2: 217–37
- O'Loughlin, K 1997. Test-taker performance on direct and semi-direct versions of the oral interaction module. In G Brindley and G Wigglesworth (eds). *access: Issues in language test design and delivery*. Sydney: National Centre for English Language Teaching and Research, Macquarie University, 117–46
- Pienemann, M, M Johnston and G Brindley 1988. 'Constructing an acquisition-based procedure for second language assessment'. *Studies in Second Language Acquisition*, 10, 2: 217–43
- Pollitt, A 1991. Response to Charles Alderson's paper: 'Bands and scores'. In J C Alderson and B North (eds). *Language testing in the 1990s*. London: Macmillan, 87–94
- Porter, D 1991. Affective factors in language testing. In J C Alderson and B North (eds). *Language testing in the 1990s*. London: Macmillan, 32–40
- Quinn, T J 1993. 'The competency movement, applied linguistics and language testing: Some reflections and suggestions for a possible research agenda'. *Melbourne Papers in Language Testing*, 2, 2: 55–87
- Quinn, T J and T F McNamara 1987. Review of Australian second language proficiency ratings. In J C Alderson, K Krahnke and C W Stansfield (eds). *Reviews of English language proficiency tests*. Washington, DC: TESOL, 7–9
- Ross, S 1992. 'Accommodative questions in oral proficiency interviews'. *Language Testing*, 9, 2: 173–86
- Ross, S 1997. May. 'CSWE outcomes: 15 issues of description and inference'. Paper presented at NCELTR National Forum on Assessment and Reporting in the AMEP
- Ross, S 1998. *Measuring gain in language programs: Theory and research*. Sydney: National Centre for English Language Teaching and Research, Macquarie University

- Royal Society of Arts (RSA) 1987. *Practical skills profile scheme*. London: Royal Society of Arts
- Royal Society of Arts (RSA) 1988. *English as a second language — dual certification*. London: Royal Society of Arts
- Selinker, L and D Douglas 1985. 'Wrestling with context in interlanguage theory'. *Applied Linguistics*, 6: 190–204
- Shohamy, E 1992. New modes of assessment: the connection between testing and learning. In E Shohamy and R Walton (eds). *Language Assessment for Feedback: Testing and Other Strategies*. Dubuque, Iowa: Kendall Hunt Publishing Company, 7–28
- Shohamy, E 1995. Performance assessment in language testing. In W Grabe (ed). *Annual Review of Applied Linguistics*, 15. Cambridge: Cambridge University Press, 188–211
- Shohamy, E 1996. Competence and performance in language testing. In G Brown, K Malmkjaer and J Williams (eds). *Performance and competence in second language acquisition*. Cambridge: Cambridge University Press, 138–51
- Skehan, P 1984. 'Issues in the testing of English for specific purposes'. *Language Testing*, 1, 2: 202–20
- Stevenson, D K 1981. Beyond faith and face validity: The multi-trait multi-method matrix and the convergent and discriminant validity of oral proficiency tests. In A S Palmer, P J M Groot and G A Trosper (eds). *The construct validation of tests of communicative competence*. Washington, DC: TESOL, 149–65
- Tarone, E 1988. *Variation in interlanguage*. London: Edward Arnold
- Tarone, E and G Yule 1989. *Focus on the learner*. Oxford: Oxford University Press
- University of Cambridge Local Examinations Syndicate (UCLES)/The British Council/IDP Education Australia 1998. *IELTS annual review 1997/98*. University of Cambridge Local Examinations Syndicate
- van Lier, L 1989. 'Reeling, writhing, fainting and stretching in coils: oral proficiency interviews as conversation'. *TESOL Quarterly*, 23, 3: 489–508

- Vaughan, C 1991. Holistic assessment: What goes on in the raters' minds? In L Hamp-Lyons (ed). *Assessing Second Language Writing in Academic Contexts*. Norwood, NJ: Ablex, 111–26
- Wapshere, D 1997. Texts and tasks: Investigating reading comprehension skills in English as a second language. In B Clayton and R House (eds). *Working away at CBA: Improving assessment practice*. Canberra: Australian National Training Authority, 101–9
- Weir, C J 1993. *Understanding and developing language tests*. Hemel Hempstead: Prentice Hall
- Wigglesworth, G 1997a. 'Task variation in oral interaction tests: Increasing the reality'. *Prospect*, 12, 1: 35–49
- Wigglesworth, G 1997b. 'An investigation of planning time and proficiency level on oral test discourse'. *Language Testing*, 14, 1: 85–106
- Wigglesworth, G and K O'Loughlin 1993. 'An investigation into the comparability of direct and semi-direct versions of an oral interaction test in English'. *Melbourne Papers in Language Testing*, 2, 1: 1–24
- Wylie, E and D Ingram 1992. 'Rating according to the ASLPR'. Unpublished manuscript, Griffith University

2

Comparing AMEP assessments: A content analysis of three reading assessment procedures¹

Geoff Brindley

Introduction

With the increasing worldwide demands for different types of assessment for certification, selection and accountability in language programs, there has been a growing desire on the part of educational authorities to be able to compare the results of different tests or assessments. There are various circumstances in which information on test comparability may be sought. For example, university authorities might want to compare the scores obtained on different tests of English for Academic Purposes, such as the Test of English as a Foreign Language (TOEFL) or the International English Language Testing System (IELTS) in order to make decisions about common standards for tertiary entry. Similarly, an organisation that uses proficiency rating scales for language teacher certification might be interested in how the scale levels relate to scores in standardised proficiency tests. In fact, in any situation where the results of more than one test or assessment are used to set a common standard for entry, exit or certification, test comparison is called for.

In recent years, the question of the comparability of different assessments has become a contentious issue in the context of the Australian Adult Migrant English Program (AMEP) in which three different tests or assessments have been used for purposes of placement, certification and summative assessment: the Australian Second Language Proficiency Ratings (ASLPR), the Certificates in Spoken and Written English (CSWE) and — until recently — the Australian Assessment of Communicative English Skills (**access:**) (see Chapter 1, this volume, for discussion).

The practice of assessing and reporting learner proficiency and achievement using different tools has led to some confusion amongst teachers in the AMEP,

since it is not clear to what extent the results yielded by these tools are comparable (Bottomley et al 1994). Although ASLPR and CSWE levels, for example, are treated as if they were interchangeable, there have been no empirical studies of the relationship between the two. Similarly, during the time that the access: test was used as part of the immigration selection process, equivalences between access: results and the other assessments used in the AMEP were not investigated, even though discrepancies were noted in the way that learners were classified.

Given the current emphasis in the AMEP on the importance of program outcomes (see Ross, Chapter 6, this volume), it is important that the assessment tools used in the AMEP should provide high quality information. However, in the absence of information on the relationship between the assessments which are used, it becomes difficult to evaluate the utility of the information they provide. For this reason a detailed comparative examination of the different assessments used in the AMEP seems called for.

This chapter reports the results of such a study in the form of an analysis of the content of the reading tests used in conjunction with access:, ASLPR and CSWE. Although the study does not address the issue of direct score comparison, such an analysis provides at least a partial basis for evaluating the validity of any comparisons — however broad — which are drawn between the three assessments under investigation.

Background

Test content analysis

Bachman et al (1996:125) argue that ‘it is a widespread view that content considerations are essential to the design of language tests and that content validity provides an important basis for the interpretation of test scores’. They go on to suggest (1996:126) that content validation requires close attention to both the language abilities to be measured and the nature of the tasks and items employed. The purpose of test content analysis is thus to describe the characteristics of test tasks and the particular aspects of language ability targeted by the test(s) or assessment tasks in question. In this way it becomes possible to compare different tests and to identify the extent to which they tap similar abilities and use similar methods. This is conventionally done by asking expert judges (usually applied linguists and/or experienced language teachers and/or

test developers) to rate test content using scales which attempt to operationalise the key dimensions under investigation (Bachman et al 1995b; Clapham 1996).

The Cambridge-TOEFL comparability study

In recent years, analyses of test content have been increasingly used in language test validation studies. The most widely reported of these is the comparison of the Cambridge First Certificate in English (FCE) and the Test of English as a Foreign Language (TOEFL), known as the Cambridge-TOEFL comparability study (CTCS) (Bachman, Kunnan, Vanniarajan and Lynch 1988; Davidson and Bachman 1990; Bachman, Davidson, Ryan and Choi 1995; Bachman, Davidson and Milanovic 1996). This study set out to determine the extent to which the two test batteries were measuring the same abilities through a detailed comparison of test content characteristics. The rating instrument used in the study was based on the framework of test method facets (TMFs) and components of communicative language ability (CLA) originally proposed by Bachman (1990) and subsequently updated by Bachman and Palmer (1996). The original rating instruments as used in the Bachman et al (1995b) study comprised a set of 12 components of communicative language ability (CLA) and 35 test method facets (TMFs), subsequently renamed ‘task characteristics’ by Bachman and Palmer (1996:47). In the light of subsequent revisions of the rating instruments, the number of TMFs has been reduced to 23 (Bachman et al 1996:131).

The CLA rating instrument used in the CTCS includes the following components from the Bachman (1990) model:

Grammatical competence	Illocutionary competence
Vocabulary	Ideational functions
Syntax	Manipulative functions
Morphology	Heuristic functions
Phonology/Graphology	Imaginative functions
Textual Competence	Sociolinguistic competence
Cohesion	Sensitivity to dialect or variety
Rhetorical organisation	Sensitivity to register

Test Method Facets (TMFs)

The set of TMFs/task characteristics used in the CTCS contains the following elements (Bachman et al 1995b:194–207; Bachman and Palmer 1996:48–57):

- *characteristics of the setting*
These include the time, place and participants involved in the test.
- *characteristics of the test rubrics*
These concern the nature of the instructions, structure of the rubrics (number of parts etc), time allotment and scoring methods.
- *characteristics of the input*
These characteristics refer to the way in which the input is presented and include the *format* of the test (eg whether the test is delivered through the aural or visual channel, whether it is delivered ‘live’, the item types used, etc); the *language of the input* (characteristics of the language itself including vocabulary, lexis, and pragmatic characteristics as well as the topics contained).
- *characteristics of the expected response*
This category is intended to describe the way in which test takers are expected to respond to test tasks or items and includes the *format* (channel, form, language etc) as described under characteristics of the input, the *type of response* required (selected versus constructed) and the *degree of speededness* (how long the test taker has to answer).
- *relationship between input and response*
This category covers *reactivity* (the relationship between the input or the response and subsequent input or responses which may vary according to whether language use is reciprocal [as in a face-to-face oral test], non-reciprocal [as in a reading test] or interactive [as in an oral interview where the interviewer varies the task to suit the level of an interviewee]). It also involves the *amount of input* that must be processed for the desired response to be given (eg how much of a reading text needs to be processed in order to respond to a given item) as well as the *directness of the relationship* (the degree to which test takers have to rely on the language in the input in order to respond as opposed to using information about the context or their own background knowledge).

The framework of TMFs/task characteristics and CLA components used by Bachman et al (1995b) has been adopted in a number of other studies of test comparability. Clapham (1993, 1996) used a modified version of the CTCS

rating scheme to investigate the content specificity of reading passages used in the International English Language Testing System (IELTS). Bachman’s (1990) TMFs have also been used as a basis for a descriptive evaluation of three tests used to evaluate the speaking ability of teaching assistants in US universities (Hoejke and Linnell 1994).

Selecting a rating scheme

The fact that the Bachman et al (1995b) rating scheme had been used in a number of previous research studies and had been modified over a period of time suggested that it would be an appropriate starting point for the analysis of test comparability in the context of the AMEP. A workshop was therefore organised during which the guidelines used by Bachman et al (1995b) were distributed, omitting those parts which were specifically related to EAP testing and thus not relevant to the present study. The rating instruments were explained and the CLA components and TMFs were discussed and illustrated at some length. Judges then attempted to apply the rating instruments to examples of the assessment tasks under investigation. However, group members found some of the categories extremely difficult to interpret and they were unable to reach agreement on the meaning of a number of the CLA components and TMFs, even after extended discussion and exemplification. Interpretation of some aspects of the CLA scales such as those associated with illocutionary competence proved particularly problematic, as did rather complex TMFs such as *cohesion* and *rhetorical organisation*.

The practicalities of using the rating instruments also emerged as a major problem. Judges were daunted by the sheer amount of time involved in evaluating 14 tasks and 85 items in terms of 12 CLA components and 23 TMFs, echoing the concern expressed by the raters in Clapham’s study (1996:148).

In the face of these difficulties, it was decided to abandon the Bachman et al scheme and to try to find an instrument which was less time-consuming to apply and which judges found easier to interpret and use.

The ALTE Framework

The scheme for the comparative analysis of test content which was finally adopted was a modified version of the descriptive checklists developed by the Association of Language Testers in Europe (ALTE nd). Although they contain many of the elements of the Bachman et al scheme which are outlined above,

the checklists are couched in somewhat more transparent language and raters reported that they found them easier to apply than the CTCS scheme. However, a major disadvantage of the checklists is that, unlike the CTCS rating instruments, there have been no published studies of their use and no information is therefore available on their validity and reliability.

According to ALTE the checklists were ‘designed for use as a practical tool for describing, in terms of a general impression, ONE VERSION of a language examination’. (ALTE nd:2, emphasis in original). They comprise a comprehensive list of test content characteristics which can be applied to different language abilities (Reading, Writing, Listening, Speaking, Structural Competence). Additional checklists are available which can be used to describe or evaluate a single task. As well as covering task and item characteristics, each checklist contains sections which allow for a general description of the test (including timing and weightings of different parts) and its presentation and layout.

The checklist used to describe the reading assessments covers the following aspects of test content, many of which are similar to the CLA components and TMFs in the Bachman et al (1995b) rating instrument:

- 1 Input (including number, type and topic of texts, target readership, writer’s intention)
- 2 Language ability tested (including main focus of testing, item type used)
- 3 Tasks (including numbers of tasks and items per task, aspects of the language used in the task, placement and sequencing of items)
- 4 Expected response (including type and range of response expected, background knowledge assumed, relationship between item and response)
- 5 Marking (including type of marking and criteria used).

Another section of the ALTE checklists concerns guidance to candidates. It covers the relevance and adequacy of candidate instructions, rubrics etc. However, this section was considered to be more relevant to standardised public tests rather than to less formal procedures such as ASLPR and CSWE (which do not contain extended guidance to candidates) and therefore could only be filled in by raters in relation to *access*. It will therefore not be reported in the analysis of the results.

The judges

Five judges participated in the study. They were all experienced ESL/EFL teachers

and/or test developers with post-graduate qualifications in TESOL or applied linguistics who had taught in Australia, New Zealand, Britain, Ireland, Canada and Finland. All but one had been involved with item development for large-scale English language proficiency tests (either *access*: or IELTS) and so were familiar with issues of task and item design and evaluation. However, none had written any of the items under consideration in the actual test comparison. All were familiar with the purpose and function of the ASLPR and CSWE.

The rating process

The ALTE checklist for reading texts and tasks was distributed to the group of judges along with the set of guidelines for users which defines and exemplifies the terminology used in the various rating categories. The judges then used the checklist to independently rate samples of test material taken from the assessments to be evaluated. Following this, ratings of content characteristics were then compared and discussed. Group members were asked to signal instances where the definitions of the task or item characteristics to be rated were unclear or where the scale categories were causing confusion. These examples were discussed and the group then attempted to reach a common understanding of the characteristic in question, based on the definitions given in the ALTE materials. As a result of these discussions, a number of changes were made to the rating scheme. These included deletions of items which were found to be confusing, superfluous or repetitious. Some items were added from the Bachman et al (1995b) scheme relating to the degree of contextual support available to the learner, the amount of new information in the text and the degree of grammatical complexity of the tasks and items. Modifications were also made to the item’s ‘main focus of testing’ which required judges to match a single skill to a particular item. These will be discussed in greater detail further on in the section describing the results of the analysis of the language skill ratings.

After these modifications had been carried out, a follow-up session was held in which judges used the revised rating scheme to rate samples of the assessments under investigation.

Content analysis of ‘non-test’ procedures: Issues and problems

All of the documented comparative analyses of test content which have been conducted to date have been carried out on standardised language proficiency tests for large populations. Particular forms of the tests or ‘papers’ under investigation have therefore been the object of comparison. For example, Bachman

et al (1995b) compared three pairs of tests: Cambridge FCE Paper 1 and TOEFL Section 3 (reading comprehension); FCE Paper 3 and TOEFL Section 2 (structure and vocabulary); and FCE Paper 4 and TOEFL Section 1 (listening comprehension).

In the present study, however, the situation was somewhat different, in that two of the three assessments being compared, the ASLPR and CSWE, unlike **access:** or **step:**, do not come with a standardised set of texts and tasks which are given to all learners. Rather, they consist of various kinds of elicitation tasks that are selected by the teachers who administer the assessment. The results are then mapped on to a set of external standards in the form of rating scale descriptors or competency statements to yield a level or statement of competency achievement.

As noted in Chapter 1, in the case of the ASLPR there are no standardised tasks or mandatory assessment materials which accompany the scale. The ASLPR contains a column entitled 'examples of specific tasks' which illustrates the types of language behaviour which might be observed at a particular level (eg *can fill out most forms regularly encountered in everyday life*). In order to elicit an appropriate sample of language use, teachers either develop their assessment materials or select from an available bank of existing materials. To this end, most institutions which use the ASLPR have assembled kits containing examples of different texts and tasks which can be used to elicit the kinds of language performances described in the scale. There is, however, no such thing as an 'ASLPR test' which is used uniformly across all AMEP teaching centres in Australia. There is a scale, there are suggestions for elicitation techniques and there are sets of stimulus materials in the form of written or spoken texts (in the case of reading and listening assessment) or oral or written prompts (in the case of speaking and writing) which vary across teachers and centres.

Like the ASLPR, the CSWE does not include nationally standardised assessment tasks, although the first edition of the CSWE (Hagan et al 1993) was accompanied by assessment guidelines containing sample tasks for different ability levels and competencies (Burrows 1993a, 1993b, 1993c). The second and third editions include examples of assessment tasks and 'benchmark' performances in speaking and writing illustrating the types of texts which learners could be expected to produce in order to meet the performance criteria for each competency (NSW AMES 1995). An 'evidence guide' giving examples of

assessment tasks is also included in each competency description. In addition, an assessment and moderation kit to assist teachers in evaluating assessment tasks and constructing their own assessment materials has been developed (Christie and Delaruelle 1997). At the time this study was conducted, however, this material was not available.

Identifying tests/tasks for the content analysis

The fact that both ASLPR ratings and CSWE competency achievement are awarded on the basis of a range of teacher-developed tasks rather than a single test makes content comparisons problematic. Clearly it is impossible to make salient comparisons between different assessments unless either a uniform set of tasks is used or, at the very least, a set of 'anchor' tasks with known measurement properties is developed which can then be used as a basis for comparison with other new tasks. Owing to constraints of time and resources, the latter course of action was not possible, so in order to address this potential dilemma, it was decided to identify a particular set of tasks which could be considered representative of each of the three assessments. This was done as follows.

I The **access:** Reading test

In the case of **access:**, the content analysis was based on the reading tasks used to assess reading as part of an exercise to set standards for the **step:** test. This was a composite version, consisting of passages and questions from several different previously used forms whose properties were already known. The test was made up of five passages and contained 40 items. The passages cannot be included for reasons of copyright. However, a general description of the tasks, texts and items can be found in Table 1.

Table 1: access: Reading test

Task#	Topic	#Items	Item type
1	Rescue of a solo sailor	10	Short answer
2a	Aboriginal rock carvings	4	Multiple-choice
2b	As above	5	True/false/not in text
3a	Kangaroo products	5	Multiple-choice
3b	As above	5	Matching
4a	Life expectancy	5	Multiple-choice
4b	As above	6	Matching

2 ASLPR reading tasks

A set of three reading tasks were selected from eight reading passages and questions aimed at Levels 1, 1+ and 2 which had been developed for use in ASLPR assessments by experienced ESL teachers from Adult Migrant Education Services, Victoria. (The tasks are included at Appendix 2.) This organisation has been using the ASLPR since the late 1970s and has devoted considerable resources to ASLPR training and development of assessment materials (McIntyre 1995). To this extent, the materials could be considered to reflect the kinds of tasks typically used to elicit performances in the context of ASLPR assessment.

It could be argued that the use of a restricted set of test activities in this way violates one of the principles of ASLPR assessment, that is, that the tasks should be selected in the light of the learner's perceived needs and interests (Ingram 1990:52). However, taken to its logical extreme, this would have meant that each learner could be given a different set of assessment tasks, making any kind of content comparison impossible. As a compromise, it was decided to constrain the number of tasks used, although teachers could still make a choice of tasks according to the perceived level of ability of the learners. In this case, given that the learners who participated in the study were enrolled at Level III of the CSWE, it was assumed that the learners' level of ability would fall between ASLPR 1 and 2 (the CSWE gives the ASLPR range of Certificate III as between ASLPR 1+ and 2, but an additional task aimed at Level 1 was included to enable 'level checking' as recommended in the ASLPR interview). The ASLPR guidelines give no precise guidance to interviewers on how many reading tasks are to be used, since this may vary according to the ability of the candidate. However, it seemed reasonable to assume that three tasks would constitute a realistic number, given the suggested time limit of thirty minutes for speaking, listening and reading (Wylie and Ingram 1992:1). The analysis accordingly focused on three tasks, corresponding to one task that might be used to assess ability at ASLPR Level 1, Level 1+ and Level 2. The tasks are summarised in Table 2. The reading passages and questions are included at Appendix 2.

Table 2: ASLPR reading tasks

Task#	Topic	#Items	Item type
1	Life in Australia	4	Short answer: oral response
2	Request to attend employment interview	5	Short answer: oral response
3	House fire	5	Short answer: oral response

3 CSWE reading assessment tasks

The CSWE reading tasks used as a basis for the content comparison were taken from the 'benchmark' texts for Certificate III provided by the developers of the CSWE (NSW AMES 1995) and thus were considered representative of the types of tasks which would be recommended for assessment of reading at this level. The tasks, which are included at Appendix 3, consisted of four parallel forms of a text and series of questions used to assess the competency *can read information texts* within three different strands of the CSWE — Community Access, Vocational and Further Study. Although CSWE reading assessment at Certificate III is aimed at assessing learners' ability to understand a variety of text types, including diagrammatic texts, procedural texts, persuasive texts and reports, it was decided to focus on this particular text type since it appeared to be the most amenable to direct comparison with those used in access: and ASLPR. However, the fact that only one text type was used means that the results of the section of the content analysis which deals with 'writer's overall intention' cannot be considered an accurate reflection of the full range of text types used in CSWE reading assessments which tap a variety of other genres.

A summary of the CSWE tasks is included in Table 3. The texts and questions are included at Appendix 3.

Table 3: CSWE reading assessment tasks

Task#	Topic	#Items	Item type
1	Coping with stress at work	6*	Short answer
2	Job profile of a clerk	6*	Short answer
3	Distance learning options in Technical and Further Education	6*	Short answer
4	Migraine headaches	6	Short answer

* Includes two-part items

Procedure

Each of the judges was given a set of rating materials for each assessment and asked to rate each passage and each item section for each test or set of tasks according to the categories outlined above. Most of the scales used required judges to rate the characteristic in question on a scale from 1 to 4, with 1 representing the 'unmarked' end of the scale. Thus a rating of 1 on the category of 'ambiguity', for example, would mean that the item was very clear while a

rating of 4 would signify that it was highly ambiguous. Means and standard deviations were derived for all ratings on each task and item using the SYSTAT for Macintosh statistics package (Wilkinson, Hill and Vang 1992).

Data analysis

General description

A general descriptive comparison of the three assessments was first undertaken according to a set of categories from the ALTE scheme which describe the numbers of tasks, items, and time allocation for the tests. These categories did not require judgments. This is reported in Table 4. Other descriptive information relating to the qualities of passages and tasks is reported further on in the section dealing with passage-based input.

Table 4: General description of test content

	access:	ASLPR	CSWE
No. of tasks	7*	3	4
No. of items	40	14	31**
Suggested time for tasks (mins)	60	15	N/A

* Two-part tasks counted separately

** Two-part items counted separately

Content ratings

Raters were asked to rate the content of the three assessments under a number of headings as follows.

Passage-based input

This part of the analysis related to the nature of the text itself, its origin, the writer's intention, topic and the type of response required.

Passage-based analysis: Tasks

The first part of this section related to the propositional content of the test in a similar way to the CTCS scheme (Bachman et al 1995b:200–202). It required judges to rate the language used in the test tasks along the following dimensions:

- degree of contextual support provided
- abstractness of language of text

- appropriacy of language in text for level of target group
- frequency of vocabulary
- degree of specialisation of language in text
- amount of cultural content
- ambiguity of language of text
- amount of new information which cannot be predicted from text
- grammatical complexity
- formality of language of text
- amount of figurative language
- appropriacy of rubric to level of target group
- appropriacy of language of items for level of target group.

Language skills tested

Main focus of testing

Raters were asked to specify the main skill or skills being tested by the item in order of importance. Provision was made for up to three skills per item to be identified. As stated previously this was to allow for the fact that an item may be testing several skills simultaneously.

Item type used

This required identification of the item format used (multiple choice, gap filling etc).

Item-based analysis: Expected response

Raters were asked to rate each item in each of the three assessments along the following dimensions relating to the language of the expected response:

- ambiguity of items
- abstractness of information in item
- appropriacy of language of item
- frequency of vocabulary
- degree of specialisation of language
- ambiguity of language of item
- grammatical complexity of language of item

- formality of language of item
- range of acceptable responses
- amount of cultural knowledge assumed.

Rater agreement

Three different indices of rater agreement were calculated. The first of these was percentage agreement between raters (RAP) as used by Bachman et al 1995b. This statistic simply reports in the form of a percentage the number of raters who agree on a given rating category. The second statistic is the weighted rater agreement percentage (WRAP) which takes into account the number of steps by which raters differ and weights the percentage accordingly (Clapham 1996:151). Thus if four raters out of five agree and the maximum difference between two raters is one step on the scale, then the WRAP would be .8. However, if they disagree by two steps, then it would be reduced to .6. Mean RAPs and WRAPs for both passage-based and item-based analysis are reported in Tables 5 and 6.

There appear to be different views in the language testing literature on what constitutes an acceptable level of percentage agreement between raters. Bachman et al (1995b) report a RAP of only 64% on the CLA ratings and 75% for the TMFs in the CTCS. However, they argue that when the RAP statistics are taken together with the G-study variance components (which were fairly low for the TMFs but somewhat higher for the CLA ratings), that this represents 'a fairly high level of agreement' (p 134). They do not appear to consider the relatively modest percentage agreement on the CLA ratings serious enough to omit these ratings from the comparative analysis of test content, even though for some CLA components the standard deviation is considerably larger than the mean rating, suggesting a very wide range of interpretations of the component in question. On the other hand, Clapham (1996:151–2) is somewhat more cautious. She reports that 'it seems that where facets have a mean WRAP of less than .7 there is too little agreement for us to believe that the raters were using the same criteria for their judgements'. In a similar vein, Thornton (1996) in a pilot validation study of the Speaking Proficiency Test (SPT) used with the Interagency Language Roundtable (ILR) Speaking Skill Level Descriptions as used by various US government agencies, set a benchmark of 70% agreement among raters. She comments that 'this bound is quite conservative and, if reached, should allow the SPT to be considered fairly

reliable' (p 33). RAPs and WRAPs of over .7 are therefore marked with an asterisk in the Tables so as to highlight those characteristics on which a minimum 70% level of agreement was obtained.

In order to obtain another estimate of agreement, generalisability coefficients (G-coefficients) were calculated for the combined ratings on both the passage-based and item-based analyses. The G-coefficient is analogous to the classical internal consistency reliability estimate. Coefficients are reported in Table 7.

Table 5: Passage-based analysis: Tasks. Rater percentage agreement

	access:		ASLPR		CSWE	
	RAP	WRAP	RAP	WRAP	RAP	WRAP
Degree of contextual support	.70*	.46	.65	.35	.50	.20
Abstractness of information	.51	.30	.53	.33	.55	.35
Appropriacy of language in text	.82*	.82*	.58	.50	.75*	.69
Frequency of vocabulary	.73*	.73*	.80*	.80*	.68	.65
Degree of specialisation of language	.59	.51	.80*	.80*	.63	.55
Amount of cultural content	.62	.51	.73*	.73*	.65	.45
Degree of specialisation of topic	.80*	.73*	.73*	.67	.75*	.65
Ambiguity of language of text	.87*	.87*	.67	.60	.70*	.70*
Amount of new information in text	.69	.61	.80*	.80*	.80*	.80*
Grammatical complexity	.54	.33	.63	.55	.60	.55
Formality of language of text	.50	.39	.60	.60	.70*	.70*
Amount of figurative language	.69	.69	.80*	.73*	.70*	.50
Appropriacy of rubric to level	.68	.68	–	–	.79*	.79*
Appropriacy of language of items for level of candidates	.79*	.79*	.33	.17	.69	.64

Table 6: Item-based analysis: Rater percentage agreement

	access:		ASLPR		CSWE	
	RAP	WRAP	RAP	WRAP	RAP	WRAP
Ambiguity of item	.53	.33	.61	.46	.78*	.73*
Abstractness of information	.64	.51	.77*	.71*	.72*	.72*
Appropriacy of language in text	.66	.59	.62	.62	.75	.74*
Frequency of vocabulary	.62	.55	.87*	.87*	.70*	.68
Degree of specialisation of language	.60	.49	.90*	.86*	.69	.68
Ambiguity of language of item	.53	.35	.64	.49	.80*	.77*
Grammatical complexity	.70*	.63	.77*	.73*	.74*	.70*
Formality of language of item	.57	.48	.79*	.76*	.69	.62
Range of acceptable responses	.95*	.95*	.57	.34	.87*	.79*
Amount of cultural knowledge assumed	.51	.17	.67	.47	.52	.34

Table 7: Generalisability coefficients: Total ratings

	access:	ASLPR	CSWE
Passage-based analysis	.75	.89	.78
Item-based analysis	.72	.61	.65

Discussion of rater agreement

It can be seen that rater percentage agreement and weighted agreement statistics are low to modest for both the passage-based and item-based analysis. In the former case, only 33 out of the 84 agreement (39%) coefficients reached .7, while in the latter case 24 out of 60 (40%) attained this threshold. With the exception of the ASLPR passage-based analysis, with a coefficient of .89, the generalisability coefficients were also modest.

Rater agreement was lowest on the following task characteristics in the passage-based analysis:

- degree of contextual support provided
- abstractness of information in text
- degree of specialisation of language of text

- amount of cultural content present in text
- grammatical complexity of text
- appropriacy of the language of the items for the level of the candidates.

Item-based analysis

The item-based analysis revealed the lowest levels of agreement on:

- ambiguity of items
- appropriacy of language used in item for level of testees
- degree of specialisation of language used in the item
- ambiguity of language used in the item
- formality of the language used in the item
- amount of cultural knowledge assumed by the item.

Here the characteristic of *amount of cultural knowledge* elicited the lowest level of agreement, with none of the agreement coefficients reaching the threshold.

Looking at the two analyses, it can be seen that there were a number of common areas where raters failed to agree. These related to the *appropriacy of the language* of the text or item for the level of the testees, the *degree of specialisation* of language of the text or item, the *formality* of the language of the text or item and the *amount of cultural knowledge* assumed by the text or item.

Low agreement on some of these categories is perhaps not surprising, however. It is very difficult for native speakers to put themselves in the position of the test-taker and to make judgments on 'appropriacy'. A question such as 'to what degree is the language of the text/item at the appropriate level?' is open to multiple interpretations. (Moreover, some studies have shown that native speaker judgments of item difficulty in tests have been at odds with the test results [cf Hamp-Lyons and Mathias 1994].) In relation to the issues of specialisation and cultural knowledge, Clapham (1996:152), in a study of the topic-specificity of EAP texts, also found low agreement among raters on topic specificity in texts. She attributes this to a lack of conceptual clarity in this characteristic and inadequate explanation. Similarly, the question regarding the extent to which the language of the text or the topic is 'specialised' raises the question of what constitutes 'specialisation'. Clearly background knowledge varies between individuals and one person's everyday knowledge may be quite esoteric for another. Here Clapham (op cit) also found lack of agreement

on this category and speculates that it may be harder to achieve agreement on specialisation than on other facets. This is clearly another category that needs a good deal of further research.

The amount of cultural knowledge required to respond to particular items or understand a given passage is also difficult to judge. Here the different backgrounds of the raters may have contributed to some divergence in the ratings, since some were more familiar with Australian culture than others. However, it is equally possible that this characteristic was particularly susceptible to misinterpretation. On closer examination, the source of the very low agreement coefficients turned out to be localised in passages 2 and 3 in the *access*: test, which concerned in the first instance the discovery of rock carvings by Australian Aboriginals and, in the second, the use of products made from kangaroos. These passages yielded ratings at extreme opposite points of the scale. Since percentage agreement indices (especially the WRAP) are very sensitive to range, this resulted in very low overall RAP and WRAP statistics for this characteristic. There may well be an explanation for this lack of agreement, however. According to the definition derived from the Bachman et al guidelines (Clapham 1996:284), 'cultural content relates to national (general) culture such as national habits, customs and beliefs'. Some judges may have perceived the passages by virtue of their overtly Australian subject matter (Aboriginals and kangaroos) as high in 'cultural content' almost by definition and therefore assumed that they required considerable cultural knowledge to interpret. On the other hand, on closer reading, other judges may have decided that despite their ostensibly heavy Australian focus, the passages were nevertheless sufficiently contextualised to be able to be understood by a person unfamiliar with Australian culture. The possibility of these two differing interpretations points to the need for extensive discussion and exemplification of the meaning of the task characteristics before using the rating instrument and for further refinement and explanation of categories such as this.

It is not clear why the criterion of *degree of contextual support* would have been problematic for raters although it may have been difficult to rate as far as the ASLPR and CSWE are concerned, since the context is partially provided in the case of the former through the interview and in the latter by the teacher. In the case of *abstractness of information* and *grammatical complexity of text*, these categories are highly subjective and may have been under-exemplified during the training session. It is interesting to note, however, that raters were able to reach reasonable levels of agreement on the latter characteristic in the item-based

analysis. This suggests that raters may find it easier to rate the complexity of specific items than it is to give a more impressionistic rating of an entire passage.

Results: Task and item characteristics

In order to compare the results of the ratings for each of the three assessments, means and standard deviations were calculated for the ratings given to each of the task characteristics in the passage-based analysis and to each of the item characteristics in the item-based analysis. Results are shown in Tables 8 and 9. In deciding whether a difference between ratings could be considered to be 'salient', the criterion used by Bachman et al (1995b:105) was adopted, that is, 'a mean difference that was larger than the standard deviation of either form being compared'. Salient differences are marked with an asterisk. A summary of salient differences showing the magnitude and direction of the differences is provided in Table 10.

Ratings of task and item characteristics

Expected response

Table 8: Passage-based analysis: Tasks

	access:		ASLPR		CSWE	
	Mean	SD	Mean	SD	Mean	SD
Degree of contextual support	3.30	1.07	3.13	1.20	3.20	.84
Abstractness of information	2.08	.82	1.87	.93	1.85	.83
Appropriacy of language in text	2.23	.40	2.25	.63	2.13	.45
Frequency of vocabulary*	2.70	.51	2.25	.63	2.13	.45
Degree of specialisation of language in text*	2.70	.62	1.80	.33	2.40	.60
Amount of cultural content	2.69	.58	2.27	.68	2.25	.74
Degree of specialisation of topic*	2.56	.51	1.87	.34	2.60	.58
Ambiguity of language of text	1.99	.22	1.87	.58	1.60	.41
Amount of new information in text*	2.89	.68	2.07	.34	2.35	.36
Grammatical complexity	3.03	.86	2.20	.61	2.60	.63
Formality of language of text*	3.03	.72	2.13	.55	2.45	.50
Amount of figurative language*	2.20	.51	1.27	.45	1.55	.90
Appropriacy of rubric to level	1.42	.57	–	–	1.30	.40
Appropriacy of language of items	2.04	.32	2.00	1.08	1.79	.56

Table 9: Item-based analysis

Task characteristic	access:		ASLPR		CSWE	
	Mean	SD	Mean	SD	Mean	SD
Ambiguity of items	1.92	1.20	1.83	.88	1.28	.46
Abstractness of information in item	1.89	.99	1.30	.58	1.35	.43
Appropriacy of language in item*	2.19	.73	1.39	.54	1.73	.43
Frequency of vocabulary*	2.13	.77	1.13	.21	1.45	.51
Degree of specialisation of language*	2.04	.82	1.14	.26	1.48	.51
Ambiguity of language of item	1.99	.99	1.67	.85	1.27	.42
Grammatical complexity*	1.89	.79	1.31	.36	1.38	.48
Formality of language of item*	2.25	.88	1.26	.41	1.75	.57
Range of acceptable responses*	1.08	.11	1.97	.97	1.25	.37
Amount of cultural knowledge assumed	2.45	1.56	1.59	.87	1.95	.82

Table 10: Salient differences between assessments

Passages		
Task characteristic	Difference magnitude	Difference direction
Frequency of vocabulary*	.57	ACCESS>CSWE
Degree of specialisation of language in text	.90/.60	ACCESS>ASPLR, CSWE>ASPLR
Degree of specialisation of topic	.69/.73	ACCESS>ASPLR, CSWE>ASPLR
Amount of new information in text	.82/.54	ACCESS>ASPLR, ACCESS>CSWE
Grammatical complexity	.83	ACCESS>ASPLR
Formality of language of text	.90/.58	ACCESS>ASPLR, ACCESS>CSWE
Amount of figurative language	.93/.65	ACCESS>ASPLR, ACCESS>CSWE
Item-based analysis	Difference magnitude	Difference direction
Appropriacy of language in item	.80/.46	ACCESS>ASPLR, ACCESS>CSWE
Frequency of vocabulary	1.0/.68/.32	ACCESS>CSWE>ASPLR
Degree of specialisation of language in item	.90/.56/.34	ACCESS>CSWE>ASPLR
Grammatical complexity	.58	ACCESS>ASPLR
Formality of language of item	.99	ACCESS>ASPLR
Range of acceptable responses	.89/.72/.17	ASPLR>CSWE>ACCESS

Language skills tested

In order to establish the degree of involvement of different reading skills in all the assessment tasks, the ALTE materials which describe and give examples of each of the skills were first distributed and discussed. A trial rating of one of the *access:* tasks was then carried out, followed by a discussion of the way in which raters had interpreted and applied the skill categories. Here a number of questions were raised concerning the feasibility of attributing particular skills to items, a concern raised in a number of research studies into the nature of reading and listening skills (Alderson and Lukmani 1989; Alderson 1990a; Buck 1990). The difficulty of aligning single skills to items is also acknowledged by ALTE (nd:3) who state in the guidelines accompanying the checklist 'that there is a great deal of ambiguity in this area'. A number of raters made the point that an item may be tapping multiple skills simultaneously. It was pointed out that a skill such as *demonstrating understanding of text as a whole*, for example, appears to depend on a number of the other contributory skills. Here some clarification was also required to distinguish between potentially overlapping categories such as *retrieving specific information* and *locating and selecting relevant information*.

In an attempt to allow for the fact that a particular item may be tapping more than one skill, therefore, the rating scheme was modified so as to allow raters to designate the degree to which they considered the skill in question to be involved by designating whether it was a *primary*, *secondary* or *tertiary* focus. Ratings were then carried out on every item for all of the assessments.

In order to obtain an overall comparative picture of the extent to which the assessments were tapping the different skills, the total number of ratings of 1, 2 and 3 given by the five raters for each skill were summed across all tasks. The number of ratings given to each skill in each of the primary, secondary and tertiary categories was then calculated as a percentage of the total ratings given for that category. Results are shown in Table 11 below which can be interpreted as follows. Taking the CSWE as an example, of all of the ratings of 1 awarded to CSWE items, 54% were awarded to *locating and selecting relevant information*, 33% to *retrieving specific information*, 3% to *demonstrating understanding of text structure*, 3% to *deducing meaning from context*, and 6% to *making inferences from information*. This method, though somewhat crude, has the advantage of allowing raters' overall perceptions of the concentration of skills to be visually displayed at the aggregate level. The number and

concentration of ratings for each skill category in each test can be seen reasonably clearly.

Given that raters were not being asked to match specific skills to items, but rather to indicate a range of skills which they thought were required, conventional inter-rater agreement indices were not calculated. Nevertheless, Table 11 which provides an overview of all ratings awarded to each task in the three assessments, gives a reasonably clear indication of the extent to which ratings clustered together for each task across the three assessment procedures.

Table 11: Language skills tested: Total ratings for each assessment

	access:			ASLPR			CSWE		
	1	2	3	1	2	3	1	2	3
Skimming for overall gist	0.5	–	5	6	5	11	–	–	57
Demonstrating understanding of text	0.5	–	–	–	8	11	–	9	–
Identifying topic of text	–	–	–	12	2	12	–	–	–
Identifying functions of text	–	–	–	3	–	–	–	–	–
Distinguishing main points	2	4	10	–	13	–	1	2	–
Retrieving specific information	38	16	36	49	38	–	33	72	–
Locating and selecting relevant information	20	22	2	25	13	–	54	12	–
Demonstrating understanding of text structure	2	–	–	–	10	22	3	–	–
Distinguishing fact from opinion	–	–	–	–	–	–	–	–	14
Deducing meaning from context	16	26	19	4	3	11	3	–	–
Making inferences from information	13	14	21	1	8	11	6	–	–
Making use of clues	2	9	2	–	–	22	–	5	29
Other	6	9	5	–	–	–	–	–	–

Passage-based input: Tasks

The section of the rating scheme entitled *passage-based input* included categories relating to the number of tasks and items, their length, the type of texts used, the purpose of each text (*writer's intention*) and the topic. Raters were also asked to identify the nature of the required response and to indicate placement of the items. Results are reported in Table 12. It should be noted here that raters frequently attributed more than one purpose to a particular text. There was also considerable overlap in the category of topic, since a *topic*

could simultaneously fall into two or more categories (for example, does a newspaper story about the closure of a hospital come under the heading of *health* or *current affairs*?). Under the category of *writer's overall intention*, therefore, in Table 12, the figures simply represent the number of tasks out of the total number which were deemed by at least two raters to be performing the function indicated. These are given rather than percentages which would be misleading with such small numbers of tasks. Thus, *provide information* was seen as the writer's intention by at least two raters on all of the **access:** and CSWE tasks, *describe* was seen as a function of one of the seven **access:** tasks, one ASLPR task and one CSWE task, and so on. The same principle is used to report the categories of *target reader* and *topic*.

Table 12: Passage-based input: Tasks

	access:	ASLPR	CSWE
# Words (mean)	546	167	375
# Items	40	14	31
Type of text			
Newspaper/magazine article	4/4	1/3	1/4
Bureaucratic document	–	1/3	2/4
Other	–	1/3	1/4
Writer's overall intention			
Provide information	7/7	2/3	4/4
Explain	0/7	0/3	0/4
Describe	1/7	1/3	1/4
Narrate	1/7	2/3	0/4
Persuade/convince	2/7	0/3	0/4
Argue for/against	0/7	0/3	0/4
Instruct/teach	0/7	0/3	3/4
Express feelings	0/7	0/3	0/4
Other	0/7	1/3	0/4
Target Reader			
General public	7/7	1/3	4/4
L2 learners	0/7	1/3	0/4
Specialist group	1/7	1/3	2/4

Table 12: Passage-based input: Tasks (continued)

Topic of text	access:	ASLPR	CSWE
House and home	0/7	1/3	0/4
Environment	2/7	0/3	0/4
Daily life	0/7	2/3	1/4
Travel	1/7	0/3	0/4
Relations with other people	0/7	1/3	0/4
Health and body care	4/7	0/3	2/4
Education	0/7	0/3	1/4
Science and scholarship	4/7	0/3	0/4
Current affairs	5/7	1/3	0/4
Food and drink	0/7	1/3	0/4
Services	0/7	1/3	0/4
Language	0/7	1/3	0/4
Other	1/7	3/3	1/4
Response			
Type of response indicated	7/7	0/3	0/4
Length of response indicated	7/7	0/3	0/4
Items sequenced as in text	2/7	2/3	4/4
Items sequenced in a different order	5/7	1/3	0/4
Placement of items			
Before the text	0/7	0/3	0/4
After the text	7/7	3/3	4/4

Limitations of the analysis

For a range of reasons, the results of the analysis of test content reported here need to be interpreted with a good deal of caution. First, as pointed out previously, the choice of ASLPR and CSWE reading texts was problematic because of the lack of standardisation of reading tasks. An attempt was made to select tasks which represented the 'state of the art' in reading task design. However, in order to establish to what degree these texts are representative of a wider universe of reading assessment tasks, it would have been necessary to carry out a wider survey of the types of texts used by teachers nationally. Unfortunately this was beyond the resources available for this study.

Second, the effects of using the ALTE rating scheme must be considered. At the time it was used in this study, the scheme had not been validated and — unlike the CTCS rating scheme — no studies of its use were available. Although an effort was made to ensure that the rating categories were discussed thoroughly with raters, the scheme is quite complex and it would have required considerable time and resources to arrive at a point where judges were confident in its use (a point also made by Clapham 1996 in relation to the CTCS content analysis materials). As with any scheme of this kind, the validity and reliability of the ALTE scheme need to be established by piloting — preferably over a considerable period of time — in order to allow the full range of interpretations of the rating categories to emerge and to be discussed. (In this context it is worth noting that various versions of the Bachman et al (1995b) scheme have been in use for nearly ten years and according to Bachman et al [1996:147] the CLA components are still in need of refinement 'both in terms of how they are defined and in the training of raters'.) The fact that low levels of rater agreement were obtained on a number of the test content characteristics suggests that the judges were not interpreting these characteristics in the same way and it is difficult therefore to have confidence in the meaningfulness of some of the content ratings.

Discussion

These limitations notwithstanding, a number of tentative observations can nevertheless be made concerning some key similarities and differences between the three assessments which emerged from the comparison.

Overall differences

Passage-based analysis

The salient differences which emerged from the ratings, along with the descriptive analysis, suggest that the assessments differ along a number of dimensions. First, in terms of the amount of material presented to learners, both the access: reading passages (average 546 words) and those used with the CSWE assessments (average 375 words) are considerably longer than the ASLPR passages (average 167 words). Second, the *frequency of vocabulary rating* suggests that raters perceive access: to contain less common vocabulary than CSWE. Third, judges rated access: much higher than ASLPR in terms of *degree of specialisation of language in text*, and *degree of specialisation of language of topic*. (However, there was low agreement on this characteristic for both the access:

and CSWE ratings, so extreme caution should be exercised in interpreting this perceived difference.) There was also a salient difference between CSWE and ASLPR on both of these dimensions, with raters showing high levels of agreement on the (low) degree of specialisation of ASLPR tasks and items. This may reflect an emphasis in the ASLPR on using texts which are familiar to candidates. The ASLPR 2 Reading descriptors, for example, refer to a reader who is able to:

... read for pleasure simply structured prose and literary and other texts which do not assume significant cultural knowledge, ability to handle complex discourse structure, or a specialist register.

Ingram (1990:50) proposes that sensitivity to specialist register becomes a key parameter only from ASLPR 3. The low *degree of specialisation* ratings here for ASLPR as compared to the CSWE passages therefore suggests that the latter may be more demanding than those used with the ASLPR.

Another difference between the assessments was seen along the dimension of *amount of new information in text*. Here the **access:** tasks were considered to contain more new information for the candidate in relation to both ASLPR and CSWE. They were also rated as containing more formal language than either of the other two procedures, as well as a greater amount of figurative language. Ratings also showed a salient difference between **access:** and ASLPR in terms of the relative grammatical complexity of the items, with the former being seen as more complex.

Item-based analysis

At the item level, there was a salient difference between **access:** and ASLPR as well as between **access:** and CSWE in terms of *appropriacy of language in item*, with **access:** receiving higher ratings. This may indicate some concerns about whether the language of the **access:** items was pitched at the appropriate level for the candidature. As with the passage-based analysis, the **access:** test was also seen to tap less frequent vocabulary and more specialised language than either the ASLPR or CSWE tasks. There were also salient differences between **access:** and ASLPR in terms of *grammatical complexity* and *formality of language of item* where once again the language of **access:** was seen as more formal and complex. Salient differences between CSWE and ASLPR were also seen in the higher ratings given to CSWE on frequency of vocabulary and *degree of specialisation of language in item*. ASLPR received the highest rating

on *range of acceptable responses*, indicating that it is the most 'open-ended' format of the three.

Nature of the input

In terms of the nature of the input, the main function of all of the reading passages was seen to be the provision of information. *Description* was also identified as a function of one passage in each of the three assessments, while *narration* figured in two of the three ASLPR passages. **access:** was the only assessment seen to contain a persuasive text. However, it should be recalled here that the selection of CSWE passages used in the study did not represent the full range of genres which learners would encounter: these would normally include a narrative text, a procedural text and a persuasive text.

Target audience

In terms of target readers, the general public were seen to be the main readership for the **access:** and CSWE passages. However, the three ASLPR passages were seen to be aimed at different audiences — the first for second language learners, the second for a specialist group (in this case, job-seekers since the text involved was a request to attend an employment interview) and the third for the general public.

Topics

It is difficult to reach any clear conclusions regarding similarities or differences between the assessments in terms of the topics sampled because of the overlap between topic areas. This meant that a number of the passages were assigned to more than one topic. However, it is perhaps worth noting that a preponderance of the **access:** tasks and half the CSWE tasks were seen to be concerned with health and body care. Science and scholarship are topics covered by **access:** but not by either of the other tests. Most of the topics covered in the ASLPR tasks appear to relate to daily life. This is not unexpected, given the focus of the ASLPR on everyday proficiency.

Response modes

Response modes differ across the assessments. As a standardised test administered under examination conditions, **access:** contains detailed rubrics for each task including the type and length of response required. ASLPR and CSWE which are administered under much more informal conditions, do not contain detailed instructions to candidates as part of the assessment task. The CSWE

has a standard rubric: 'read the text and answer the questions' and the class teacher explains how to do the task, while the ASLPR reading assessment relies on verbal instructions given by the interviewer.

Item types

Another difference between the assessments was in the diversity of item types used. Whereas **access:** and ASLPR use a single item format — the short answer question (in oral mode in the case of the ASLPR) **access:** contains a variety of item types — short answers, true/false/not given, multiple-choice, and matching.

Skills tested

The analysis of the skills ratings suggests there is a relatively heavy clustering of ratings in skills 6 (*retrieving specific information*) and 7 (*locating and selecting relevant information*) on almost all of the tasks in each assessment. With the exception of skills 10 and 11, *deducing meaning from context* and *making inferences from information*, which appear consistently in the ratings for the **access:** tasks, and *identifying topic of text* and *skimming for overall gist* in the ASLPR ratings, the pattern of ratings suggests that raters considered few of the other skills to be engaged in the assessment tasks.

Overall, the ratings suggest that the reading passages used with ASLPR and CSWE sample a more restricted range of skills than those used in the **access:** test. Only the **access:** test was deemed to be assessing 'higher order' skills such as *deducing meaning from context* and *making inferences from information* to any significant degree.

Summary

This chapter set out to examine similarities and differences in the content of the reading tasks which are used in the assessment procedures under investigation. This involved the subjective rating of the characteristics of all reading passages and items by a group of five expert judges using a rating instrument requiring multiple ratings of test content characteristics. Although it was originally intended to use the Bachman et al (1995b) scheme which had been used in previous comparative studies of test content, this rating instrument did not prove to be usable by the raters. This necessitated the adoption of a modified version of an unpiloted content rating scheme developed by the Association of Language Testers in Europe (ALTE). This scheme employs rating categories

which are partially based on the Bachman et al scheme, but was found by raters to be more transparent and relevant to the population at hand.

Despite the training of raters in the interpretation and application of the rating instrument, there was quite low inter-judge agreement on a number of the content characteristics. Differences between test content characteristics which could not be consistently rated should consequently be interpreted with caution.

The results of the content rating exercise revealed a number of salient differences between the assessments. On the task ratings, the **access:** test received higher ratings than the ASLPR on dimensions related to the *degree of specialisation of language in text*, *degree of specialisation of topic*, *amount of new information in text*, *formality of language of text* and *amount of figurative language*. Salient differences were also found between **access:** and CSWE on *frequency of vocabulary*, *amount of new information in text*, *formality of language of text* and *amount of figurative language*, and between CSWE and ASLPR on the two *specialisation* criteria, with ASLPR receiving the lower rating. At the item level, differences were found between the three assessments in terms of *frequency of vocabulary* and *degree of specialisation of language in text*. Here there were salient differences between **access:** and CSWE and between CSWE and ASLPR. ASLPR was judged as permitting the most open-ended responses. The **access:** test was perceived by raters to be tapping a wider range of skills than the other two assessments.

The ratings identify **access:** as the most demanding assessment of the three. This finding is not particularly surprising, given that **access:** is a standardised test designed to assess levels of ability ranging from lower intermediate up to quite advanced. What is perhaps of more interest in the context of the AMEP curriculum are the differences between ASLPR and CSWE. In addition to differences in text length, the language used in CSWE tasks and the topics of the texts were seen to be more specialised, the grammar more complex, the amount of figurative language greater, the vocabulary less frequent and the language of the items more formal. Considered in combination, these results suggest that the ASLPR assessment tasks are considered to be somewhat less demanding than those used with the CSWE III assessments and that it would therefore be easier for a learner to gain a rating of ASLPR 2 than to succeed on a CSWE Certificate III task. However, in order to seek further evidence for this hypothesis, it would be necessary to compare scores of the same group of

learners on both assessments, an original intention of this study which was not possible owing to an incomplete set of data on CSWE competency attainment.

Conclusion

This study revealed a number of key differences between the three assessment procedures under investigation. In the process, however, it raised a number of serious methodological questions concerning the viability of test content comparisons which use ratings of expert judges. In the first place, such comparisons are based on the assumption that native speaker judges are able to rate the test characteristics from the point of view of the test taker. When a judgment concerning the difficulty, complexity or familiarity of a task or item is called for, native speakers are required to estimate the difficulty for the population of test takers for whom the test was designed. (The rating instrument used in the CTCS, for example, asks judges to rate rubrics for 'prepared' and 'unprepared' test takers [Bachman et al 1995:197] and to rate the propositional content of passages and items 'in relation to the specific group of test takers for whom the test is intended' [Clapham 1996:283].) Such an assumption is difficult to sustain, however, since native and non-native speakers will not necessarily perceive the test items or approach the test tasks in the same way (cf Alderson 1995:122). A related problem affecting the reliability of the rating instruments used is the highly individual nature of test-taking processes. Judgments of aspects of test content, such as the perceived clarity of an item or of the extent to which it taps background knowledge, are extremely subjective and may vary across test-takers, whether they be native or non-native speakers, thus making it difficult to achieve consensus among raters. A further problem affecting rater consistency relates to the feasibility of asking judges to match specific language skills to particular items. Here a number of research studies have found that judges are unable to assign specific language subskills to items with any degree of consistency (Alderson and Lukmani 1989; Buck 1991). Moreover, some researchers have argued that in the case of reading and listening test items which involve interactive and parallel processing, it may not be possible to match particular skills to particular items since any given item may be tapping several skills simultaneously (Alderson 1990a; 1990b; Brindley 1997).

The difficulties experienced by raters in arriving at common interpretation of the content characteristics may have been partially due to the raters' lack of experience in using the scheme at hand. On the other hand, it is interesting to note that the results of other test content analyses have identified similar diffi-

culties. In her study of the topic specificity of reading passages used in IELTS, an international test of English for Academic Purposes (EAP), Clapham (1996), for example, found quite high levels of consensus on some facets, such as 'Grammar' and 'Cohesion' but low agreement on others such as 'Cultural References', 'Contextualisation–Cultural Content' and 'Contextualisation–Topic Specificity'. She reports (1996:150) that:

None of the three raters were confident about their assessments. They felt that although some facets were unambiguous and straightforward to answer ... they were still worried by others. They all said that they would not expect to give the same ratings another time as they felt their internal rules for assessing the facets kept changing.

Clapham (op cit p 153) attributes this uncertainty to inadequate explanation of some of the facets and a lack of definitional clarity. In a similar vein, Bachman et al (1996:135), commenting on the modest levels of percentage agreement (in the order of 64%) between raters on the CLA components in the CTCS, refer to 'ambiguities in definitions of the characteristics and how these are understood by raters'.

This is not to suggest, however, that achieving consensus between raters is an impossibility. Bachman et al (1995b, 1996), using the variance components from a generalisability study as an index of judge consistency, found very low variance components for judges, indicating 'virtually perfect agreement among raters' on the TMF ratings (p 104). Lumley (1993) also reports substantial agreement among judges on the subskills tested by particular items in an EAP reading test and on item difficulty. Other studies involving expert ratings have also reported acceptable levels of agreement following training (Weir et al 1990; Anderson et al 1991). These findings suggest that expert judges may be able to rate aspects of test content consistently provided adequate opportunities to use and discuss the rating instruments are provided.

Despite the problems experienced in this and other studies, test content analysis nevertheless remains one of the few ways in which it is possible to systematically describe and compare the characteristics of different tests. However, many of the problems identified in this study — particularly those surrounding the definition of content characteristics and the highly subjective nature of test-taking processes — are far from being resolved. As researchers who have carried out detailed comparative studies of test content acknowledge, further refinement and clarification of the rating instruments used for test content analysis

need to be undertaken (Bachman et al 1996:135; Clapham 1996:162). At the same time, in order to complement data gathered from native speaker judges, a good deal of research will be required into the perceptions of test content of subjects who represent the actual target population, using both introspective and retrospective methods (eg Buck 1990; Alderson 1990b). Such research will constitute a vital and necessary part of the ongoing validation of any instruments used for analysis of test content.

Acknowledgment

1 I would like to thank John Langille for his assistance over a considerable period of time in assembling the content rating instruments, coordinating the rating process and processing the data from the rating exercise reported in this chapter. Sincere thanks are also to Erin Beggs, Brent Merrylees, Clare McDowell and Valerie Roantree who gave up their time to rate thousands of items of test content.

References

- Alderson, J C 1990a. 'Testing reading comprehension skills (Part One)'. *Reading in a Foreign Language*, 6, 2: 425–38
- Alderson, J C 1990b. 'Testing reading comprehension skills (Part Two)'. *Reading in a Foreign Language*, 7,1: 465–503
- Alderson, J C 1995. 'Responses and replies'. *Language Testing*, 12, 1: 121–4
- Alderson, J C and Y Lukmani 1989. 'Cognition and reading: Cognitive levels as embodied in test questions'. *Reading in a Foreign Language*, 5, 2: 253–70
- Anderson, N, L Bachman, K Perkins and A Cohen 1991. 'An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources'. *Language Testing*, 8, 1: 41–66
- Association of Language Testers in Europe (ALTE) 1996. *European language examinations*. Cambridge: University of Cambridge Local Examinations Syndicate
- Association of Language Testers in Europe (ALTE) Undated. *Development and descriptive checklist for tasks and examinations*. Cambridge: University of Cambridge Local Examinations Syndicate
- Bachman, L F 1990. *Fundamental considerations in language testing*. Oxford: Oxford University Press
- Bachman, L F and A S Palmer 1996. *Language testing in practice*. Oxford: Oxford University Press
- Bachman, L F, A Kunnan, S Vanniarajan and B Lynch 1988. 'Task and ability analysis as a basis for examining content and construct comparability in two EFL proficiency test batteries'. *Language Testing*, 5, 2: 128–59
- Bachman, L F, B K Lynch and M Mason 1995a. 'Investigating variability in tasks and rater judgements in a performance test of foreign language speaking'. *Language Testing*, 12, 2: 238–57
- Bachman, L F, F Davidson, K Ryan and I-C Choi 1995b. *An investigation into the comparability of two tests of English as a foreign language*. Cambridge: Cambridge University Press
- Bachman, L F, F Davidson and M Milanovic 1996. 'The use of test method characteristics in the content analysis and design of EFL proficiency tests'. *Language Testing*, 13, 2: 125–50
- Bottomley, Y, J Dalton and C Corbel 1994. *From proficiency to competencies: A collaborative approach to curriculum innovation*. Sydney: National Centre for English Language Teaching and Research, Macquarie University
- Brindley, G 1997. 'Investigating second language listening ability: Listening skills and item difficulty'. In G Brindley and G Wigglesworth (eds). *access: Issues in language test design and delivery*. Sydney: National Centre for English Language Teaching and Research, Macquarie University, 65–86
- Buck, G 1990. *The testing of second language listening comprehension*. PhD Thesis, University of Lancaster, 76–77
- Buck, G 1991. March. 'Expert estimates of item characteristics'. Paper presented at *Language Testing Research Colloquium*, Princeton, New Jersey
- Burrows, C 1993a. *Assessment guidelines for the Certificate in Spoken and Written English. Stage 1*. Sydney: NSW Adult Migrant English Service
- Burrows, C 1993b. *Assessment guidelines for the Certificate in Spoken and Written English. Stage 2*. Sydney: NSW Adult Migrant English Service
- Burrows, C 1993c. *Assessment guidelines for the Certificate in Spoken and Written English. Stage 3*. Sydney: NSW Adult Migrant English Service
- Christie, J and S Delaruelle 1997. *Assessment and moderation: Book 1. Task*

- design*. Sydney: National Centre for English Language Teaching and Research, Macquarie University
- Clapham, C M 1993. Is ESP testing justified? In D Douglas and C Chapelle. (eds). *A new decade of language testing research: Selected papers from the 1990 language testing research colloquium*, 257–71
- Clapham, C 1996. *The development of IELTS: A study of the effect of background knowledge on reading comprehension*. Cambridge: Cambridge University Press
- Davidson, F and L Bachman 1990. The Cambridge–TOEFL comparability study: An example of the cross-cultural comparison of language tests. In J H A L de Jong (ed). *Standardization in language testing*. AILA Review, 7: 24–45
- Hagan, P, S Hood, E Jackson, M Jones, H Joyce and M Manidis 1993. *Certificate in Spoken and Written English*. Sydney: NSW Adult Migrant English Service and the National Centre for English Language Teaching and Research, Macquarie University
- Hamp-Lyons, L and S P Mathias 1994. 'Examining expert judgements of task difficulty on essay tasks'. *Journal of Second Language Writing*, 3, 1: 49–68
- Hoefke, B and K Linnell 1994. "Authenticity" in language testing: Evaluating spoken language tests for international teaching assistants'. *TESOL Quarterly*, 28, 1: 103–26
- Ingram, D E 1990. The Australian Second Language Proficiency Ratings (ASLPR). In J H A L de Jong (ed). *Standardization in language testing*. Amsterdam: Free University Press, 46–61
- Lumley, T J N 1993. 'The notion of subskills in reading comprehension tests: An EAP example'. *Language Testing*, 10, 3: 211–34
- McIntyre, P 1995. Language assessment and real-life: The ASLPR revisited. In G Brindley (ed). *Language assessment in action*. Sydney: National Centre for English Language Teaching and Research, Macquarie University, 113–44
- New South Wales Adult Migrant English Service (NSW AMES) 1995. *Certificates in Spoken and Written English*. Sydney: NSW Adult Migrant English Service

- Thornton, J 1996. *The unified language testing plan: Speaking proficiency test. Spanish and English pilot validation studies*. Arlington, Virginia: Center for the Advancement of Language Learning
- Weir, C J, A Hughes and D Porter 1990. 'Reading skills: Hierarchies, implicational relationships and identifiability'. *Reading in a Foreign Language*, 7, 1: 505–10
- Wilkinson, L, M Hill and E Vang 1992. *Statistics. Systat for Macintosh*. Evanston, Illinois: SYSTAT, Inc
- Wylie, E and D Ingram 1992. 'Rating according to the ASLPR'. Unpublished manuscript

3

Issues in the development of oral tasks for competency-based assessments of second language performance

Gillian Wigglesworth

Introduction

In recent years, there has been a growing interest among second language acquisition researchers and language testers in the cognitive and contextual factors which affect second language task performance. A major focus of investigation has been the influence of different task types on both the quality and quantity of linguistic output (see, for example, Robinson 1996, forthcoming; Wigglesworth 1997; Skehan 1998b). One of the findings which has consistently emerged from this line of research is that language performance may vary along a number of dimensions according to the type of elicitation task which is used and the conditions under which it is implemented (Douglas 1994; Fulcher 1996; Tarone 1998).

These findings are of considerable importance in the context of language assessment, since they suggest that variations in task characteristics and task conditions can make a task easier or more difficult for language learners. This potential variability in task difficulty is of particular concern in assessment and reporting systems which use different assessments devised by individual teachers as a basis for reporting learners' achievement against predetermined standards or benchmarks. If learners are given tasks of unequal difficulty and their performances are then mapped on to the same standards, it could be argued that learners who undertake the more difficult tasks are disadvantaged. It thus becomes important to identify the key factors that contribute to task difficulty so that these can be taken into account, and, as far as possible, controlled, in assessment task design.

This chapter reports on the outcomes of a research project which set out to investigate the manner in which variations in task conditions and task characteristics influence the assessment of spoken language competencies in the context of the Certificates in Spoken and Written English (CSWE) in Australia (see Brindley, Chapter 1, this volume). One of the main aims of the project was to provide a research base for the implementation of the CSWE by providing empirically-developed guidelines which could assist teachers to design competency-based assessments. At the same time, it was hoped that the findings would be applicable to teaching practice by providing information on the kinds of tasks that could be used not only for assessment purposes but also for classroom teaching.

The effects of task-related variables on oral language production

Researchers have investigated a range of variables which may affect second language speaking performance. These include the planning time available to the speaker, the cognitive load imposed by the task and whether the interlocutor is a native speaker (henceforth NS) or non-native speaker (henceforth NNS). A number of factors may influence cognitive load. For example, chronological sequencing reduces cognitive demand, whereas multiple actions and actors increase it (Candlin 1987). Such features have provided a starting point for the operationalisation of cognitive difficulty. Empirical investigation has suggested that cognitive load, or difficulty, does influence performance, although not always adversely. In this regard, Robinson et al (1995) found that a more cognitively demanding task in fact produced higher accuracy rates in the use of articles and higher levels of lexical density. Similarly, Wigglesworth (1997) found that more accurate language was produced by high intermediate level students under test conditions where the task was more cognitively demanding and was identified by statistical analysis as more difficult.

Investigation of the differences between NS and NNS interlocutors has shown that negotiation is more likely in NNS/NNS interaction than in NS/NNS interaction (Varonis and Gass 1985). Plough and Gass (1993) suggest that the less a shared background exists between conversational partners (linguistically and culturally) the greater the frequency and complexity of negotiation routines. Plough and Gass (1993) also investigated the effect of interlocutor familiarity. They analysed the discourse using a number of measures related to the cooper-

ative nature of the discourse, and found that while interlocutor familiarity had little effect on the use of back channel cues, it affected sentence completions, interruptions and overlaps. Interviewer familiarity also affected the way comprehensibility and misunderstanding was dealt with, with a higher incidence of comprehension checks and clarification requests occurring in the familiar dyads. Where the interlocutors were unfamiliar, echoic repetitions of the interlocutor speech appeared to be used to avoid potential misunderstandings. Where there was a lack of familiarity it resulted in greater numbers of interruptions showing greater involvement and commitment, but where tasks were familiar, there was greater negotiation in the form of confirmation checks and clarification requests. Overall, however, the interlocutor variable had a greater influence on the language output than did task familiarity.

Planning time has been the focus of a number of studies. In these studies, measures of accuracy, fluency and complexity have been used to analyse the discourse. Ellis (1987) found greater accuracy in the use of the past tense where planning time was provided. This result was in contrast to a study by Crookes (1989) who found no effect for accuracy with planning time. Wigglesworth (1997), however, found a variable effect, with planning time appearing to influence accuracy only for high proficiency candidates on more cognitively demanding tasks. Foster (1996) found a trend toward increasingly accurate language where planning time was available, but the differences were not significant. These somewhat variable results have been reflected in a series of studies in which both the type (detailed versus non-detailed) and amount (length of time in minutes) of planning time were examined with three different task types (Foster and Skehan 1996; Skehan 1996; Skehan and Foster 1997). The analyses undertaken in this series of studies suggested that time and task type influence different aspects of language. For example, a narrative task elicited more complex language at the expense of accuracy, while a personal exchange task elicited more accurate language (based on error-free clauses) but not more complex language. This trade-off effect between accuracy and complexity may result from differences in the perceived goals of the task (Skehan 1996). Skehan argues that the provision of structure and planning time are resources the second language learner may draw on to reduce the processing load on a limited capacity processor.

While these studies have all employed detailed analyses of the discourse output, comparisons across different studies must be conservative. Generally, there has been little systematicity in the analytical approaches adopted, and the critical

constructs (usually *accuracy*, *fluency* and *complexity*) have been operationalised differently by researchers. It is therefore difficult to summarise and draw conclusions, although we may tentatively conclude that there is a complex interaction between task type, cognitive load and planning time. However, in order to make significant comparisons across different studies, a well-defined unit for the analysis of language data is needed (Foster et al 1998).

The research findings outlined above are of clear relevance to assessment situations. Assessment tasks aim to elicit an adequate sample of language in order to make an appropriate assessment of the testee's ability. However, if the characteristics of the tasks and the conditions under which they are administered influence the nature of the response, this may result in samples of language which are variable in terms of the key features which serve as assessment criteria, such as grammatical accuracy, syntactic complexity and fluency. In the assessment situation, the extent and the degree of influence of such variables on task output therefore need to be clearly identified so as to ensure that the tasks present a comparable level of challenge.

In language assessment, the test developer's goal is to minimise as far as possible all sources of measurement error which are external to the learner's language performance. This is done in order to ensure that the score the learner obtains is, to the greatest extent possible, a true reflection of his or her ability to use the language for the purposes required by the assessment and not a product of influences unrelated to the ability being assessed, that is, what Messick (1989) calls 'construct-irrelevant variance'.

There are many sources of error which are external to the learner. For example, in an oral interview, the interviewer may vary in the manner in which the interview is conducted and this may affect the learner's ability to demonstrate his or her range of linguistic knowledge. The nature of the task may also impact upon the final score obtained by the learner in that it may be a more or less difficult task. The person who is rating the performance (who may or may not be the interviewer) will also influence the score since one rater may be more or less lenient than another. In addition to these factors, there are a variety of affective variables and personal attributes which may influence the learner's performance (Bachman 1990).

Task-based assessment offers a way of ascertaining how well the language learner can use the language for communicative purposes. We noted that a

wide range of factors may contribute to variability in language task performance. A learner's score, or rating is thus a complex product of multiple influences. This study focuses on the issues related to the tasks. Tasks may vary in terms of either the characteristics inherent in the task (features internal to the task such as the amount of structure provided, cognitive load or familiarity of the content) and the conditions under which they are administered (such as the availability of planning time, or whether the interviewer is a NS). Variations in task characteristics or the conditions under which they are administered may systematically influence the language output of the test candidate (Skehan 1998a). This issue was the focus of the investigation.

Methods

In this project, the findings from previous studies of the effect of task variables on language output were drawn upon and used to identify a series of key variables related to task characteristics and task conditions. These variables were then systematically manipulated and empirically examined in order to determine their effect on the performances of the learners on a range of different tasks.

All tasks were competency-based classroom assessment tasks routinely used and administered by teachers in order to evaluate student achievement of spoken language competencies in the CSWE. The tasks were developed from a range of tasks sent in by teachers from three Australian states. Five commonly taught competencies were identified at two different CSWE levels. At Certificate II (post-beginner level)¹ competencies 5, 6 and 7 were selected while assessment at Certificate III (intermediate level) focused on competencies 5 and 6, as listed below. Brief descriptions of the assessment tasks are provided in Appendix 4.

CSWE Level II

Competency 5: Giving instructions

Competency 6: Negotiating an oral transaction to obtain information

Competency 7: Negotiating an oral transaction for goods and services

CSWE Level III

Competency 5: Obtaining information through a telephone enquiry

Competency 6: Negotiating a complex/problematic spoken exchange

Two or three tasks were developed to assess each competency. The tasks were required to conform to certain criteria as follows:

- contextual material that needed to be pre-taught was kept to a minimum
- the context of the tasks was universally familiar to learners
- the tasks needed to be limited as far as possible to the skill that was being assessed
- the tasks needed to be relevant to learner needs.

Once the tasks had been developed, one task for each competency was selected as a control task which was undertaken by all of the learners. This task was not manipulated in any way. The remaining tasks (the manipulated tasks) were manipulated using specifically identified variables in either an independent two-way or four-way design.

Identification of variables

Based on an analysis of the research literature on factors affecting second language task performance, two task characteristics and two task conditions were identified as variables for manipulation. The task characteristics were:

- 1 *structure*: the task was developed either with or without structure. This was operationalised in terms of the amount of information provided to the learners to assist them in doing the task. Specifically, where structure was present, the learners were provided with five specific prompts to direct them in their interaction with the interlocutor. Where structure was not provided, one general statement was provided to guide the learners in the task.
- 2 *familiarity*: this was operationalised according to whether the task activity was an activity with which the learners would reasonably be expected to be familiar. However, this variable proved problematic in two important respects. First, it was not *a priori* possible to determine whether the various situations were or were not familiar to the learners. Second, although for the three other variables the tasks were identical, this was not possible with the familiarity variable since in these tasks it was the situation which changed, and it was not possible to determine definitively whether the tasks were exactly equivalent in difficulty prior to the collection of data.

The task conditions were:

- 1 *NS versus NNS interlocutor*: this was operationalised according to whether the interlocutor involved in the exchange was a NS of English or a NNS.
- 2 *planning time*: planning time was operationalised as either five minutes planning or no planning time. Planning was always manipulated in conjunction with one of the task characteristics.

These variables were assigned to the tasks as shown in Figure 3.

	Task 1	Task 2	Task 3	Task 4	Task 5
Certificate II Comp 5	No variation	+ planning + familiar	- planning + familiar	+ planning - familiar	- planning - familiar
Certificate II Comp 6	No variation	+ structure	- structure	+ familiar	- familiar
Certificate II Comp 7	No variation	+ structure	- structure	NS interlocutor	NNS interlocutor
Certificate III Comp 5	No variation	- planning + structure	+ planning + structure	- planning - structure	+ planning - structure
Certificate III Comp 6	No variation	NS interlocutor	NNS interlocutor	+ structure	- structure

Figure 3: Allocation of variables by task and task type

The learners

There were 80 learners at each level drawn from different ESL centres in the three states participating in the project. Learners were representative of AMEP clients, coming from a range of language background, representing a range of ages and both genders. Tasks were administered by two or three trained and experienced teachers in each state. As shown in Table 13, all learners were administered Task 1 for each competency. In addition, each learner was randomly assigned one of the manipulated tasks for each competency. This meant that each manipulated task was undertaken by approximately 20 learners, although numbers were sometimes smaller due to occasional errors in task administration.

Table 13: Task and subject assignment

Certificate II	Task 1	Task 2	Task 3	Task 4	Task 5
Competency 5	80	20	20	19	20
Competency 6	80	20	20	20	20
Competency 7	80	20	19	21	20
Certificate III					
Competency 5	80	20	20	19	21
Competency 6	80	21	20	20	19

Tasks were administered by teachers trained and experienced in CSWE assessment. Each learner undertook four (level III) or six (level II) tasks individually over a two- or three-day period. All performances were tape recorded for rating at a later stage. Once the learners had completed the assigned tasks, they were asked to rate the difficulty of the tasks in which they had participated on a five-point Likert scale.

Rating procedure

The CSWE assessments are carried out using binary, criterion-referenced scales. However, since it was important in this project to ensure that measures of task difficulty were as sensitive as they could be to the range of variation possible within any particular feature of performance, an analytic scale was used for the rating. This scale had previously been extensively trialled and used in the assessment of the English language proficiency of adult immigrants and was familiar to raters (see O'Loughlin 1997). Each of the rating scales was accompanied by a set of descriptors covering four criteria at each of seven levels. All performances were double rated by randomly assigning the performances across 16 trained and experienced raters.

Quantitative analysis

The analysis of the data was carried out using both quantitative and qualitative approaches. The quantitative part of the study aimed to identify differences in the difficulty level of each of the oral tasks and to do this three separate quantitative evaluations were undertaken. This was necessary as the numbers of learners were relatively small and robust differences were unlikely to emerge. First, analyses of variance (for four-way comparisons) and t-tests (for two-way comparisons) were performed on the raw scores provided by the raters.

Second, the data were subjected to a Rasch analysis, using the statistical modelling program, FACETS (Linacre 1990). This program produces an estimate of candidate ability based on all the available information — that is, the different 'facets' of the assessment situation — which may be considered to impact most seriously upon the assessment environment. In this case, the information included in the analysis consisted of four facets: the candidate, the task, the rater and the rating criteria. The program uses all of this information to model an estimated ability value expressed in a unit called a *logit* for each of the candidates. In addition to the ability estimates, the analysis provides a logit value for each of the facets identified — the difficulty of the task, the relative leniency or harshness of the rater, and the relative difficulty of the criteria. Thus it is possible to compare the relative difficulty of these facets and, in this case, the difficulty of the tasks.

Although such analyses are not usually performed with the relatively small numbers used here, there are precedents for its use as a research tool with small N sizes (see, for example, Lumley et al 1994). However, as the number of candidates who undertook each of the manipulated tasks was small (approximately 20 in each group), the additional measures of task difficulty were considered essential. The third measure of task difficulty came from the learner evaluations of the difficulty of the tasks they had undertaken. This was calculated by determining the proportion of the learners who graded each task they had taken on a five-point Likert scale from very easy to very difficult.²

Results

The first analyses of the oral data investigated whether there were any significant differences in the performances of learners from the different centres by analysing the scores they achieved on Task 1 which was common to all learners. There were no significant differences for learners participating at either CSWE level, and although significant differences across the tasks were identified, no interaction effect was present in the way particular centres interacted with particular tasks.

The discussion below focuses on the manipulated tasks (Tasks 2–5) within each competency. The Certificate II tasks are discussed first.

Certificate II

Competency 5: Giving instructions

This oral interaction task involved a four-way manipulation with planning time

and familiarity. For Tasks 2 and 3, the learner was required to give instructions about how to use a bank automatic teller machine. For Tasks 4 and 5, the learner was required to explain to a 12-year-old child how to change a light globe. The manipulated variables for each task and the mean scores for the raw score analysis and Rasch analyses are presented in Table 14a.

Table 14a: Certificate II, Competency 5, raw scores and Rasch estimates

Tasks:	ATM machine instructions (familiar) Changing light globe (unfamiliar)		
	Variables	Raw scores	Rasch estimates
Task 2	+ planning/+ familiar	14.17	11.9
Task 3	- planning/+ familiar	11.87	8.6
Task 4	+ planning/- familiar	14.40	12.5
Task 5	- planning/- familiar	14.39	14.5

Where the raw scores are concerned, the higher the score, the easier the task is likely to be. Therefore, in order to ensure that the polarity of the two analyses was in the same direction, the average raw scores for each task were subtracted from 28, the total possible score (4 criteria x 7 rating points). Table 14a gives the total scores for all four criteria averaged over the sample size. Task 3 (the more familiar task — operating the ATM — without planning time) appears to be the easiest task, although this difference is not significant ($F(3,147) = 2.077$, $p = .106$). For the Rasch analysis, once again, it appears Task 3 is the easiest of the tasks, supporting the view that the unfamiliar task is the most difficult. In this analysis, planning time appears to provide a small advantage where the task is unfamiliar, but a disadvantage when the task is familiar. This latter finding is in line with previous results discussed in the literature and is discussed further below.

Table 14b: Certificate II, Competency 5, student evaluations (%)

	Easy		Average		Difficult		Total N
	N	%	N	%	N	%	
Task 2	11	61.1	6	33.3	1	5.5	18
Task 3	7	36.8	9	47.4	3	15.8	19
Task 4	6	30.0	10	50.0	4	20.0	20
Task 5	9	45.0	9	45.0	2	10.0	20

The learner evaluations (Table 14b) suggest that Task 2 (familiar with planning) was seen as the easiest of the four tasks. The three remaining tasks appear to be assessed as having similar difficulty levels by the learners. It is important to bear in mind, however, that this is a somewhat rough measure as the learners have only completed the control task and one of the manipulated tasks, not the full range of tasks.

Competency 6: Negotiating an oral transaction to obtain information

Two different variables were investigated for this competency. In Tasks 2 and 3 structure was manipulated, with planning time included in both. For Tasks 4 and 5, familiarity was manipulated (here operationalised as the learner being required to either obtain information about a language school for themselves [more familiar] or about a secondary school for their child [less familiar]).

In all four of these tasks, the specified interlocutor was a NS. Unfortunately, an administrative error meant that the majority of candidates attempted Task 5 (less familiar) with a NNS interlocutor. For this reason, it was decided to remove the scores for the few learners who had had a NS interlocutor so that the NS/NNS interlocutor was not confounded within Task 5. However, this meant that the NS/NNS variable was confounded across Tasks 4 and 5, since candidates doing Task 4 (more familiar) interacted with a NS, whereas candidates doing Task 5 (less familiar) interacted with a NNS. This position was adopted to ensure that the variables were clear in the analysis.

Table 15a suggests that there is little difference between the structured and unstructured tasks where the raw scores are concerned, but a notable difference in Tasks 4 and 5, with the less familiar task appearing to be the easier of the two. This difference is significant ($t = 2.589$ $p < .05$ $df = 33$).

Table 15a: Certificate II, Competency 6, raw scores and Rasch estimates

Tasks:	Information re art exhibition (+/- structure) Information re English classes (familiar) Information re secondary school (unfamiliar)		
	Variables	Raw scores	Rasch estimate
Task 2	+ structure	14.75	10.5
Task 3	- structure	13.73	16.3
Task 4	+ familiar	13.67	6.9
Task 5	- familiar	10.67	5.6

Examination of the Rasch logit estimates shows there is a clear difference between the structured (Task 2) and unstructured (Task 3) with the unstructured task being the more difficult. While the differences in this analysis are small for the other two tasks (Tasks 4 and 5), Task 5 (less familiar) again appears marginally easier.

The learner evaluations (Table 15b) support this view with Task 3 (unstructured) being designated as the most difficult and Task 4 (more familiar) being slightly more difficult than the less familiar task.

Table 15b: Certificate II, Competency 6, student evaluations (%)

	Easy		Average		Difficult		Total N
	N	%	N	%	N	%	
Task 2	6	30.0	11	55.0	3	15.0	20
Task 3	5	26.3	9	47.4	5	26.3	19
Task 4	12	60.0	6	30.0	2	10.0	20
Task 5	9	56.2	6	37.5	1	6.2	16

Competency 7: Negotiating an oral transaction for goods and services

In this competency, structure was manipulated in Tasks 2 and 3 and interlocutor (NS versus NNS) in Tasks 4 and 5. The NNS interlocutor in Task 5 was a NNS at a similar level of proficiency to the learner taking the task. With the structured/unstructured tasks (Tasks 2 and 3), an administrative error meant that a few candidates on these tasks interacted with NNS interlocutors rather than NS interlocutors. Consequently, these learner scores were removed from the analysis of these tasks. The raw scores suggest minimal differences here, although the unstructured task appears to be the easier of the two. Where the interlocutor variable was present (Tasks 4 and 5) the NNS appears to make the task easier.

The Rasch estimates also indicate little difference between the structured and the unstructured tasks, and agree with the raw score tallies indicating that the NNS interlocutor makes the task an easier one. However, these effects are small.

Table 16a: Certificate II, Competency 7, raw scores and Rasch estimates

Tasks:	Variables	Raw scores	Rasch estimates
	TV repair (+/- structure)		
	Newspaper delivery (NNS versus NS interlocutor)		
Task 2:	+ structure	13.56	8.5
Task 3:	- structure	12.82	8.7
Task 4:	native speaker	13.90	10.2
Task 5:	non-native speaker	12.70	6.8

The learner evaluations support the view that the presence of a NS interlocutor makes the task more difficult. However, given the small differences in the values, we need to treat this finding with some caution. The smaller numbers for Tasks 2 and 3 reflect the exclusion of learners who interacted with a NNS as discussed above.

Table 16b: Certificate II, Competency 7, student evaluations (%)

	Easy		Average		Difficult		Total N
	N	%	N	%	N	%	
Task 2	13	76.5	4	23.5	0	0.0	17
Task 3	8	53.3	6	40.0	1	6.7	15
Task 4	12	57.1	6	28.6	3	14.3	21
Task 5	10	50.0	9	45.0	1	5.0	20

Certificate III

Competency 5: Obtaining information through a telephone enquiry

For this task type, the same task was manipulated with structure and planning time. The results are shown in Table 17a and 17b.

Table 17a: Certificate III, Competency 5, raw scores and Rasch estimates

Task:	Variables	Raw scores	Rasch estimates
	Job advertisement		
Task 2:	- planning/+ structure	9.63	6.8
Task 3:	+ planning/- structure	10.37	10.6
Task 4:	- planning/- structure	10.76	12.3
Task 5:	+ planning/- structure	11.43	12.5

Table 17a suggests that task difficulty increases across the tasks. While the differences in the raw scores across tasks are not significant for the summed raw scores, the general pattern suggests that the presence of structure advantages learners, while the presence of planning time disadvantages them. This same general pattern is exhibited for the Rasch analysis and for the learner evaluations (Table 17b). Task 5 was clearly the most difficult from the students' point of view.

Table 17b: Certificate III, Competency 5, student evaluations (%)

	Easy		Average		Difficult		Total N
	N	%	N	%	N	%	
Task 2	2	10.0	15	75.0	3	15.0	20
Task 3	8	40.0	9	45.0	3	15.0	20
Task 4	5	26.3	10	52.6	4	21.1	19
Task 5	6	28.6	6	28.6	9	42.9	21

From the evidence above, for this series of tasks we may rather tentatively propose that the unstructured tasks are the most difficult and that the presence of planning time appears to increase the difficulty.

Competency 6: Negotiating a complex/problematic spoken exchange

The final series of tasks was again subdivided into two groups of two. Tasks 2 and 3 were identical except that the interlocutor in Task 2 was a NS and in Task 3 was a NNS. In Tasks 4 and 5, structure was manipulated. Planning time was present in both cases.

Table 18a: Certificate III, Competency 6, raw scores and Rasch estimates

Tasks:	Variables	Raw scores	Rasch estimates
	Negotiating annual leave (NNS versus NS interlocutor)		
	Negotiating complaint (+/- structure)		
Task 2:	native speaker	10.60	14.3
Task 3:	non-native speaker	9.60	8.0
Task 4:	+ structure	9.60	8.6
Task 5:	- structure	10.80	8.8

While the differences in the raw scores were not significant, the NNS interlocutor apparently made the task the easier of the Tasks 2 and 3. For Tasks 4 and 5, there was again a small difference in the raw scores, with the unstructured task being the more difficult. The Rasch difficulty estimates also suggest that interaction with a NS task (Task 2) is more difficult than with a NNS (Task 3). This is supported by the learner evaluations (Table 18b). More learners found Task 2 more difficult than they did Task 3. The Rasch analysis reveals no difference for Tasks 4 and 5.

Table 18b: Certificate III, Competency 6, student evaluations (%)

	Easy		Average		Difficult		Total N
	N	%	N	%	N	%	
Task 2	9	42.9	6	28.6	6	28.6	21
Task 3	8	40.0	9	45.0	3	15.0	20
Task 4	3	15.0	10	50.0	7	35.0	20
Task 5	6	31.6	7	36.8	6	31.6	19

Compared to the student evaluations of the other task types, the learner evaluations here suggest that all but Task 3 are perceived as relatively difficult. This suggests that it is task type (negotiating a complex/problematic exchange) that is difficult, with the NNS interlocutor reducing the difficulty.

These results are not conclusive in a statistical sense, however. This, no doubt, results partly from the fact that the numbers of learners taking the different variations of the tasks were small. However, we may draw some tentative conclusions based on the fact that the differences are in most cases consistently in the same direction on all three types of analyses and across the different tasks.

The effects of task type: Summary

In summary, for each of the task characteristics the findings of the quantitative analysis were as follows:

- 1 *Structure*: the weight of the evidence points toward structure making the task easier in three of the task types (task type 2; task type 4; task type 5). The most robust of these is task type 4. The exception is task type 3, where the raw scores suggest the unstructured task is slightly easier.
- 2 *Familiarity*: the results for this variable were problematic. In Certificate II,

Competency 5, where the activity was more familiar the task appeared to be easier. However, where familiarity was manipulated in Certificate II, Competency 6, the less familiar of the two tasks appeared to be easier. However, the Certificate III, Competency 6, task administration was compromised by the use of NNS interlocutors, so it is not possible to definitively tease out whether the effect is for familiarity, or for the nature of the interlocutor. This issue is discussed further below.

For task conditions we may summarise as follows:

- *NS versus NNS*: where the interlocutor is a NNS the task appears to be easier (task type 3; task type 5);
- *Planning*: planning time was manipulated in a four-way design with familiarity (task type 1) and structure (task type 4). The results tentatively suggest that a familiar activity is easier where planning time is *not* present. Further, planning time appears to adversely influence performance in both structured and unstructured tasks.

Qualitative analysis

Although quantitative analysis of the type described above can reveal general patterns of influence on task difficulty, it cannot show precisely how a specific task affects learners' output at the linguistic level. In order to do this, it is necessary to examine samples of the oral language produced by learners in response to the tasks. It was therefore decided to undertake a more detailed qualitative analysis of the oral performances. All of the performances were transcribed and a number of aspects of the learners' language were identified which were influenced by various characteristics and conditions of the task. These are discussed below in relation to each variable investigated, but first some general points pertaining to pragmatic and cultural issues which relate to task design are identified and briefly discussed.

Some tasks require a greater level of pragmatic competence than others. For example, negotiations and complaints appear to put more pragmatic demands on learners where they must make a decision about appropriate levels of directness or indirectness, how much to comply with the interlocutor's suggestions, and what degree of renegotiation is acceptable. In addition, students must find a balance in terms of politeness which relates both to the student's view of their assumed roles, and actual role (student/teacher).

Students from particular language and cultural backgrounds may find these kinds of encounters more problematic than other groups, and are potentially penalised in these kinds of assessments if these points have not been addressed in class. While further analysis and discussion of these issues was not within the scope of this study, they should prove a fruitful area for future empirical investigation.

The role of the interlocutor is also an important one. The use of NNS versus NS interlocutors is discussed in some detail below. In addition, a broader discussion is included which addresses the manner in which the interlocutor can influence, both positively or negatively, the task from the learner's point of view. The intention of the discussion is not to critique interlocutor behaviour, but to focus awareness of the kinds of issues which teachers may wish to take into account when administering tasks to students for assessment purposes.

Structure

In general, the results of the quantitative analysis suggest that learners handled the structured tasks more successfully than unstructured tasks. The dialogues tended to follow a particular *script* based on the outlines of the role play cards. While this provided less opportunity for some of the more proficient students to demonstrate a broad range of skills, it provided support for less confident students. Tasks in which structured input was provided resulted in data that was more homogeneous across the different learners than was the case with the matched unstructured task. Two samples of discourse are provided in Examples 1a and 1b below from learners doing the structured task in Certificate II, Competency 6. Although all learners are at a similar level, it is notable that in the unstructured task the learner tends to elicit a much smaller range of information about the exhibition. In fact in the unstructured task, often only one aspect (for example how to get there, cost) was asked about, compared to the five pieces of information identified in the structured task.

Example 1a: [Structured task, Certificate II, Competency 6/Task 2]

- I: Good morning, State Museum.
 S: Oh, hello, could you tell me er Aboriginal painting exhibition?
 I: Yes.
 S: Information.
 I: Yes, certainly, what would you like [to know]?
 S: [What time] opening time?

- I: Right, it um it's the same as the museum opening times which are every day from ten o'clock to five o'clock and er except Sundays where it's one o'clock to five o'clock.
- S: Thank you. Er how much /does it/ cost do you cost sorry, [how much] do you cost?
- I: [/?/] There is a small charge for the um exhibition itself um that's seven dollars fifty an adult um.
- S: Yes.
- I: And if er there are some concession fares for student concession ... tickets.
- S: Oh thank you.
- I: And that's five dollars.
- S: Oh five dollars.
- I: Stu/ student concession.
- S: OK.
- I: Ticket.
- S: How about guided tour?
- I: Yes o/ of the exhibition itself there are guided tours every half hour.
- S: Mm.
- I: Um and you can just join one of the tours um if you wish to.
- S: Oh thank you. Er could you could you have um do you have another language translation?
- I: Er yes there are um a number of other language um er information tours um about the ... well not so much tours they're personal information where you use a headset um with a cassette giving you the information in other languages. Wha/ which is your language, Madam?
- S: Yes Yes I'm Japanese.
- I: Right yes we have er we have cassettes um available with headphones um for a small charge um and you can listen to the information in your own language.
- S: Oh thank you.
- I: Er the charge on that um it's about er, I'm just checking, yes it's two dollars um to hire the cassettes.
- S: Yes thank you. When it finish Aboriginal painting exhibition?
- I: Oh right yes um well it's going another couple of months we er finish in the thirty-first of July
- S: Thirty-first of July.
- I: Yes.
- S: Thank you very much.
- I: That's OK, my pleasure I hope you en/ come and enjoy it.
- S: Yes of course thank you.
- I: Thank you.
- S: Bye.
- I: Bye.

Example 1b: [Structured task, Certificate II, Competency 6/Task 2]

- I: Good morning, State Museum.
- S: Yeah good morning. Um I want, I have /problem. I, I want, I would like to mm to see the um ... I would like to see the /Aboriginals/ er painting.
- I: Yes.
- S: Er what er what time the, what time er does it, er, open?
- I: Er the museum opens every day from ten o'clock to five o'clock except Sundays when it's um, one o'clock to five o'clock.
- S: Yeah, thank you.
- I: OK, anything else?
- S: How how /mu/ how much is this cost?
- I: There is a small charge for the er Aboriginal exhibition. That's what you wanted to see the Aboriginal paintings?
- S: Yeah.
- I: Yes um it's seven dollars fifty.
- S: Yeah.
- I: For an adult.
- S: Seven dollar fifty.
- I: Students have a small concession if you have er if there's a student [/??/]
- S: [yes I have] a concession card.
- I: Oh right OK [yes um it's] it's five dollars for a [student concession card]
- S: [yeah OK] [yeah /???/] mm Who, who with wh/ who do /?/ with with me with me to go tour?
- I: Er you/ are you a group is it just one person you want to go on a tour?
- S: Yes.
- I: Yeah, there are guided tours.
- S: Yeah.
- I: Every, every half hour um.
- S: Yeah.
- I: For the um, exhibition.
- S: Yeah, OK.
- I: OK.
- S: Mm is that information and tour available in language other than English?
- I: Um, yes. What language do you speak?
- S: Yeah, I'm er Vietnamese.
- I: Vietnamese. Yes, we have um recorded tape um tape tours and you can use the headphones um there's a small charge for that. Um, it's about two dollars to hire the headli/ headphone set and the tape recorder and then you can listen to the um information.

- S: [Yes] [Yes] I also, yeah, I also the hear the er hear listen to the English.
 I: Yes, er OK all right, you can er you can join either the I/ the English tour or you can pay a short er small fee and have your own headset. [OK]
 S: Yeah [OK]. Um when, when, is this er finish?
 I: Er the exhibition finishes um on July the thirty-first.
 S: Oh yeah, yeah OK.
 I: OK.
 S: Yeah, thank you.
 I: Anything else?
 S: No, no.
 I: All right. Thank you very much. Bye.
 S: Bye bye.

Example 2: [Unstructured task, Certificate II, Competency 6/Task 3]

- I: State Museum, can I help you?
 S: Yes, mm, I want to ask you what time you open? I want to see exhibition of Aboriginal painting.
 I: Um well the first date is the fifteenth of May and the opening hours are from nine until four in the afternoon.
 S: Uhu. Mm ... how about tomorrow I come to se/
 I: Yes you can come tomorrow.
 S: OK. How much will cost tickets?
 I: Ah, the tickets cost er ten dollars.
 S: Ten dollars. That's OK. Maybe tomorrow I will come ten o'clock.
 I: OK.
 S: OK thank you very much.
 I: OK.
 S: Bye.
 I: Bye.

Using the information from the cards to ask questions in basic form allowed less proficient students to maintain a relatively successful interaction with the interlocutor by means of the prompt alone, although in these tasks familiarity, rather than structure, was the variable manipulated. However, the prompts for both Tasks 4 and 5 contain a certain amount of structure and the examples illustrate the importance of careful consideration as to the role such a struc-

tured prompt may play. The ability of the candidate to maintain the interaction and obtain the information by reference to the prompts on the role play cards casts some doubt on the learners' comprehension of their interlocutors' extended responses and on their achievement of the particular competency.

Interlocutors' cards for tasks assessing the student's ability to obtain information were worded in such a way that the teacher often supplied the information, unsolicited by the student. In this way the purpose of the competency (to obtain information) being assessed was sometimes negated, especially in the case of the more passive or hesitant students. This makes it possible for interlocutors to change the competency being tested by adopting a particular role. For example, rather than obtain information through requests, the learner becomes a passive receiver of information which is supplied unsolicited by the interlocutor. This situation may be the result of the interlocutor's awareness of the intended content of the dialogue from role cards, as well as from the wording provided, eg the card states that the teacher should 'give the student information about x, y and z', rather than respond to the student's requests for information.

In Certificate II, Competency 7, the variable manipulated was NS versus NNS interlocutor. The interlocutor card for this task, however, included information for the interlocutor to provide to the learner. In Example 3, at several points, the teacher interlocutor supplies the language that the role card is intended to elicit.

Example 3: [Certificate II, Competency 7/Task 4]

- I: Good morning.
 S: Good morning
 I: Er Mayor's newsagency can I help you?
 S: Yes I want to er deliv/ er.
 I: You'd like to what?
 S: /reserve/
 I: Yes to receive to receive which paper Sir?
 S: Er Australian newspaper at home only on Saturday.
 I: Ah only on Saturday?
 S: Yes.
 I: OK um.
 S: Um I want to know the cost for ... /reserve/ at home The Australian newspaper.

- I: Yes certainly now when would y/ when would you like to start delivery?
this Saturday?
- S: This Saturday.
- I: This Saturday.
- S: If possible yes.
- I: OK the cost for three months is um twelve dollars just for a Saturday uhu OK um and
can you give me so you'd be happy with that so the first Satur/ Saturday the fifteenth
of June we'll start delivering your paper.
- S: Uhu.
- I: Um.
- S: And the cost is twelve dollar.
- I: Twelve dollars for three months.
- S: For three months.
- I: Yes [on special] at present.
- S: [OK] uhu.
- I: Um can you give me your name please?
- S: My name is Mr (gives name).
- I: Right and your address?
- S: Um my address is (gives address).
- I: And what is your telephone number please Mr (name)?
- S: My telephone number is (gives telephone number).
- I: (repeats telephone number) ...
- S: ...
- I: ...
- S: ... (clarification of telephone number)
- I: Um now how are you going to pay for this, will you come round to the
newsagent or do you.
- S: Yes you/ can I have a bill [every] month?
- I: [yes] yes [certainly]
- S: [if possible]
- I: Certainly we'll send you a bill every month.
- S: And at what time er receive your newspaper?
- I: Er on a Saturday morning we usually deliver them between oh round about quarter
past seven in the morning.
- S: OK.
- I: OK.
- S: It's OK.

- I: es that will be the latest time it mi/ might be earlier but it certainly won't be later than
seven fifteen am.
- S: OK.
- I: Thank you.
- S: Thank you.
- I: Goodbye.
- S: Bye.

Precise wording of tasks is important to ensure that similar language samples are elicited from each learner. For example, Certificate II, Competency 6, Tasks 2 and 3 required the learner to obtain information about an exhibition from a local art gallery. Task 2 was structured and Task 3 was not. The prompt for the unstructured task simply required students to 'obtain information about the exhibition'. This was construed as information about the *content* of the exhibition by some students and interlocutors, rather than more practical information, as was set out in the structured task. This resulted in some interlocutors inadvertently complicating the issue by supplying information about the artwork itself, about the provision of headphones, wheelchair access etc. Since the language required for these interactions is more sophisticated than that required for obtaining practical information, some learners are potentially disadvantaged. An example of the complications which could arise with the unstructured task is given below.

Example 4: [Unstructured task, Certificate II, Competency 6/Task 3]

- I: Morning, State Museum.
- S: Oh yes, er yesterday I see the newspaper and see the advertisement about the special
/exhibi/ exhibition of /ab/ Aboriginal paintings. Is this?
- I: Yes, we do have that exhibition on at the moment. It's a very interesting exhibition.
- S: Oh, yes, I'm very interesting, but I I want to know um the exhibition er what time will
open?
- I: Um well it's on throughout the opening times of the museum. The museum is open
every day from one, from sorry, I beg your pardon from um ten o'clock to five o'clock
except on Sundays when it's open from one o'clock to five o'clock.
- S: Oh, but er, how much for er one person?
- I: There is a small charge to go into the exhibition. It's seven dollars fifty an adult.
- S: Mm
- I: Um and there is a er concession er, charge for students if you have a student
concession card or um and for children which is five dollars.

- S: Oh, have you er ... ?
 I: There is also a family ticket you can get.
 S: Family ticket oh have you er um er a lunch including er coffee or tea?
 I: No there there is a shop um a caf/ a small cafeteria where you can get coffee and tea but that's not included in the price.
 S: No. But er I know I, I know er in the exhibition um what types of painting in there?
 I: Well the paintings um are all of course done by the er Aboriginal people um and many I think from the people from um Arnhem Land in the north of Australia, um the Aboriginal people in Arnhem Land.
 S: Arnhem Land.
 I: Yes.
 S: Because I, I /b/ I interested in the painting of views.
 I: /Painting of views/
 S: Have you a lot of these paintings in the exhibition?
 I: Um well the Aboriginal paintings are very um distinct. They're quite unique um and yes they they Aboriginal paintings usually tell a story but it's done in a a different way.
 S: Oh.
 I: Er from the traditional painting of a view. There are some um Aboriginal painting artists who paint er the typical um view or landscape paintings.
 S: Oh oh. This exhibition, is it er suitable for the family together to go to see it?
 I: Yes yes, er it's er a very interesting exhibition um adults and children enjoy it. Erm as I said we do have er a family ticket two adults and two children, which is about twelve dollars.
 S: [Yes it's good, is is good.]
 I: (yes twelve dollars)
 S: Is good.
 I: Yes.
 S: Oh so, I have one er one question ask you.
 I: Yes, certainly.
 S: Um the exhibition will open Fri/ er Monday to Friday Monday to Saturday except Sunday, is it?
 I: Ye, that's right yes, er it is open on Sunday as well but only in the afternoon.
 S: Oh, only in afternoon.
 I: Yes.
 S: Yes um when I when I make sure er to go to the exhibition.
 I: Yes.
 S: Is it for er is it book now or um is it booking?
 I: You don't need to book. You can just come um and er attend and go through the exhibition.

- S: Oh.
 I: At your leisure.
 S: Oh that's good, OK.
 I: Yes.
 S: Thank you.
 I: OK.
 S: OK, bye.
 I: All right. Thank you very much, bye bye.

With respect to the structured task (Task 2), an examination of the discourse revealed that certain interlocutors appeared to be trying to follow the agenda set out in the task, which was unavailable to the students. This resulted in the interlocutor taking a very dominant role and providing extended information on the topics set out in the alternative task.

Another point was raised in relation to the task used in Certificate II, Competency 7, Tasks 2 and 3, where structure was again manipulated. The functional purpose of this competency is to 'negotiate an oral transaction for goods and services'. Although some students tried to negotiate time of visit and cost of service, the analysis of the discourse revealed that in some cases relatively little *negotiation* actually took place. Instead, the majority of the interaction was taken up with describing the problem with the machine and providing names and addresses. Relatively little time was taken up with negotiating the time. In addition, it was found that this task tended to predispose a dominant position by the interlocutor. An example of this behaviour follows; Example 5 is from the structured task, and Example 6 from the unstructured task.

Example 5: [Structured task, Certificate II, Competency 7/Task 2]

- I: Good morning washing repair service can I help you?
 S: Good morning yeah er can me help you my machine machine doesn't working.
 I: Oh OK then what's the problem?
 S: I think it's er it's problem with the er er change er er change water yeah.
 I: Yes something with the water.
 S: Yeah.
 I: What is the machine doing?
 S: I think it's er for er er it's there it's water it's all /over the floor/
 I: Oh dear OK.
 S: Yeah.

- I: There's water all over the floor er right OK well now we'll need to come to look at that.
- S: Mm.
- I: Er turn the taps off would you while you the machine is is not working.
- S: Not working yeah.
- I: Have you turned the taps off?
- S: Yeah.
- I: OK good alright well let's see what's your name please?
- S: My name is Nutzet, yeah.
- I: Nutzet.
- S: Yeah.
- I: Is that your first name or your second name?
- S: Er it's my first name yeah.
- I: Yes.
- S: And my family name it's (gives name)
- I: (repeats name)
- S: Yeah (repeats and spells name)
- I: And your address Mr (name)?
- S: Er my address is er (gives address) yeah.
- I: Oh good and your phone number please.
- S: Er my phone number is er (gives phone number).
- I: Oh good OK thanks very much now um er I need to tell you now it's 40 dollars for us to visit your home to look at the machine.
- S: Mm.
- I: Is that OK?
- S: Yeah it's OK and er and you for someone to come and fix it /?/?/
- I: Yes alright er now what about Friday um ten o'clock in the morning OK?
- S: Ten o'clock in the morning.
- I: On Friday.
- S: Er it's alright yeah I not going out in the time yeah.
- I: OK F/ right this Friday ten am.
- S: Yeah.
- I: And er we will b/ come round to your house and have a look at your machine for you.
- S: Yeah alright.
- I: No worries, OK goodbye.
- S: Thank you very much.

Example 6: [Unstructured task, Certificate II, Competency 7/Task 3]

- I: Good afternoon, Dandenong washing machine repair service, can I help you?
- S: Good afternoon my washing machine doesn't work er can you help me /?/
- I: Yes I I expect so um what's the problem, Madam, what's wrong?
- S: Mm er it's no work just not work.
- I: It's it's just not working.
- S: Yes.
- I: Oh I see OK alright well turn the taps off please er so that there's no water er and could you give me your name.
- S: Yes (gives name).
- I: Yes could you please spell it please surname first if you wouldn't mind.
- S: (Spells name).
- I: (Repeats name) yes and the first name.
- S: (Spells name).
- I: (Repeats name).
- S: Yes.
- I: Yes OK Miss or Mrs (name).
- S: Miss.
- I: Miss (name) and your address please.
- S: (Gives address).
- I: (Repeats street name) yes er phone number please.
- S: (Gives phone number).
- I: Good OK er now um will you be paying cash or a cheque?
- S: Cash.
- I: Cash OK, we do take Bankcard would that be better?
- S: Mm no matter.
- I: No matter alright well cash it is then we'll um.
- S: Er um er er ... in when we can ... arrive.
- I: When, when oh let me see er Friday would be a good day will you be home on Friday morning?
- S: Yes.
- I: How about ten o'clock?
- S: Yes it's OK.
- I: OK ten o'clock at (address) good um right now I need to tell you it's 40 dollars for us to come to your house to look at the machine OK.
- S: OK.
- I: And then if there's new parts it may be more expensive alright.
- S: Yes.

- I: OK.
 S: OK I understand.
 I: Good alright Miss (name) we'll see you at ten o'clock on Friday.
 S: Thank you bye.
 I: Bye bye.

This type of exchange highlights an important aspect of the assessor's role in the CSWE, namely that the interlocutor, in interacting with the learner, needs to behave in such a way as to ensure that the particular text type required for the assessment (in terms of social and functional purpose) is elicited. A two-pronged approach to this needs to be adopted: first, tasks need to be very carefully designed in order to ensure that the social and functional purpose is paramount, and second, interlocutors may need explicit training to ensure that the appropriate task type results from the interaction.

Relating this more specifically to the structured or unstructured nature of the task, the interlocutors' cards in both the unstructured and the structured versions of the tasks are identical, identifying for the interlocutor a list of items to be discussed with the student. This means that although in the unstructured task, the student has not been provided with structured input, the interlocutor nonetheless is able to provide a strong framework for them since the interlocutor's card is to some degree structured. This is particularly the case in Certificate II, Competency 7 and Certificate III, Competency 6, where the functional purpose of the task is to demonstrate the ability to negotiate a particular situation. The quantitative analysis of Certificate II, Competency 7 suggests that, unlike the cases where structure is manipulated in the other tasks, the unstructured task is easier. As can be seen from Examples 5 and 6 above, the content of the unstructured dialogue is very similar to that of the structured dialogue, as the student responds to information-seeking questions. This tends to reinforce the view that the interlocutor is largely responsible for the structure and content of these role plays which require negotiation to take place. Note that in the two competencies which require the students to obtain information (Certificate II, Competency 6 and Certificate III, Competency 5) learners perform better where structure is present. However, in these cases, the students are required to ask the questions. In the competencies where negotiation is required, the interviewer and the student can both answer questions, and therefore it is much easier for the interviewer to consciously or unconsciously structure the interview for the students.

Further support for this view can be obtained from examination of the transcripts for Certificate III, Competency 5, a task in which the social/functional purpose was to obtain information through a telephone exchange. The task involved the manipulation of structure with planning time, and required learners to make enquiries about a job. The qualitative analysis supported the quantitative findings discussed above — that the more structured task allowed improved performance on the task overall. Structure was found to be critical in this task for constraining the content of the interactions. On the whole, examination of the discourse suggested that fluency suffered in this task with numerous long pauses and general hesitancy. In the structured task, this occurred as learners attempted to formulate the diverse range of questions required for items listed on card. However, where structure was not present (Tasks 4 and 5) students displayed greater hesitancy and an apparent lack of ideas concerning the information they should obtain, reflecting the lack of input provided. A more detailed analysis of this aspect of the discourse is required. Examination of the transcripts also suggests that grammatical errors are more frequent in the unstructured tasks. However, the detailed discourse analysis required to support these observations empirically is outside the scope of this project. Finally, the lack of structure also results in the content of the dialogues being more variable than for structured discourse. There is a more random nature to the interactions, with a lack of coherence between utterances and more ambiguities, leading to misunderstandings and the need for clarification. This allows the interlocutor to introduce complications and less relevant topics, which are more demanding for weaker students as shown in Example 4.

Familiarity

This variable was manipulated in two tasks. In Certificate II, Competency 5, the more familiar task involved a description of how to use an ATM, and the less familiar task required the learner to explain how to change a light globe. For a number of reasons, these tasks were problematic. First, some students were unfamiliar with the exact procedures and were, therefore, unable to fully complete the tasks. Second, there was some ambiguity about the imaginary settings for these particular interactions which caused confusion in several instances. For example, it was unclear whether the events were taking place at the time of speaking or the instructions were for a future scenario. This produced a range of highly variable data across learners in terms of the language functions required as shown in Examples 7a and 7b. The extracts below are

from the task which required learners to describe how to use an ATM machine. Example 7a demonstrates the long interactive role play which results from the task; Example 7b shows how differently this task may be constructed, resulting in an extremely brief monologue. These findings identify the importance of piloting all task materials on a range of students, and preferably with a range of interlocutors, before they are used for assessment purposes.

Example 7a: [Certificate II, Competency 5/Task 2]

Interactive role-play:

- I: Oh Vinh, can you help me, I don't know how to use this card.
 S: Yeah. Er, if er if you have an automatic teller machine card.
 I: No.
 S: You want you want the er push the push out the money?
 I: Yes [I want to].
 S: [in] Commonwealth Bank?
 I: Yes
 S: Er you um you can you /can/ um the start and start you can push the button, um, your PIN number.
 I: My PIN number.
 S: Your PIN number.
 I: Oh yes I [remember] my [PIN number] yes.
 S: [yeah] [PIN number] yeah [/?/]
 I: [so I]
 S: You put PIN number again.
 I: Yeah.
 S: Then you er you put OK [OK].
 I: [yes] yes.
 S: /a/ and withdraw and push the withdraw button.
 I: Um where's that withdraw, oh yes.
 S: On the on the on the left [on the left].
 I: Yes [right OK].
 S: And you push the saving or the mm /depend you/ [saving] [/?/?/ [saving] yeah.
 I: Oh [right] er I er[yeah I have a savings] account yes.
 S: You push the button saving.
 I: Savings account, yes.
 S: Yeah and then you push the, and number the money er.

- I: OK I want to take out fifty dollars.
 S: Yeah how many money and how.
 I: Fifty dollars.
 S: Yeah how much.
 I: So where do I put fifty dollars.
 S: Then yeah you push the er final you push OK you er you wait, you wait.
 I: OK OK yes and now what I wait.
 S: Yeah you wait then the machine, er, pay you the card er card and er the paper the write the mm the mo/ the money of you.
 I: Oh yeah.
 S: And er you you put the money [laughs].
 I: OK.
 S: You get money.
 I: Oh thank you, that's easy.
 S: Yeah.
 I: Thank you very much, for your help [/?/
 S: Oh no problem.
 I: Bye.

Example 7b: [Certificate II, Competency 5/Task 2]

Short monologue:

- S: Put into ... ATM card inside the machine. ... Press your /private/ number. ... Look at the machine er er look at the [/?/. Choose choose what do you want first er, ... Receive money or your card. Finish.

Without an interlocutor card, there was some ambiguity about the role of interlocutors. For example, they could become a silent listener, a passive interactant or an active interactant, as shown in the examples above. It is clear that the role of the interlocutor, as discussed in more detail below, can be crucial to the outcome of the task for the student.

Finally, this task appears to have been fairly demanding for most students. This was reflected in much hesitation, confusion, and prompting by the teacher. For some students, unfamiliarity with specific vocabulary items caused problems.

Ultimately, the language produced in response to this task varied considerably in terms of complexity and coherence.

In Certificate II, Competency 6, students were required to obtain information through an oral transaction. Familiarity was manipulated in Tasks 4 and 5. In Task 4, students were required to obtain information for themselves in a familiar environment. In Task 5, the information was required for someone else in an unfamiliar environment. The 'less familiar' task (Task 5) was confounded by the use of NNS in the majority of the sample. However, although this task was intended to be less familiar, the structured input on the prompt card was more detailed and comprehensive than for Task 4, and hence may have resulted in more successful interactions, outweighing the familiarity variable. In addition, although the concept of requesting information for a third party was considered to be more demanding than requesting information for oneself, the items on the card did not require the use of the third person on the whole. They could be constructed as straightforward 'wh-' type questions, eg How many students are in each class? etc.

Native versus non-native speaker interlocutor

The use of NS or NNS interlocutors is a variable given in the range statement of the CSWE competencies. This allows teachers or assessors to choose whether to administer the task with a NS or a NNS. In including this in the study as a variable, there were two distinct groups of NNS who emerged as interlocutors. The first group were highly proficient and competent NNS from outside the classroom. The second group were NNS who were members of the class itself. In examining the discourse for these tasks the following observations were made.

Where the NNS were highly proficient, students appeared to be more nervous than was the case with either their peers or the language teachers. Tasks were generally carried out in a straightforward manner, in terms of their scope and the language used, closely following the instructions on the cards. The interlocutors tended to adopt the more authentic role of a stranger, rather than the role of a 'language teacher' who is familiar with the communicative strategies of students.

With the NNS student group, which formed the majority of the NNS used, there was a great range in the level of proficiency of those used as interlocutors. The interactions tended to be informal and were conducted as pair work activi-

ties rather than formal assessments. Certain interlocutors failed to cover the full extent of the task, omitting critical items and curtailing the dialogue. Others were very hesitant and incoherent, leading to problems of comprehension and an apparent lack of fluency on the part of the assessed student. The student's competency was, therefore, difficult to assess in these interactions. Others, however, were more confident and proficient, giving a structure and coherence to the interactions. For this reason there was greater variability with this group of interlocutors. This needs to be taken into account when deciding whether or not to use NNS interlocutors in competency-based assessment tasks.

Planning time

Planning time was manipulated together with structure in Certificate III, Competency 5 and with familiarity in Certificate II, Competency 5. In the Certificate II tasks, planning time made no difference where the task was unfamiliar, but with the familiar task, planning time, if anything, disadvantaged learners. In Certificate III, in Tasks 2 and 3, which required learners to obtain information about a job, a structured task was provided with planning time in Task 3 and no planning time in Task 2. In Task 3, the dialogues appear to be more complex, fluent and extensive with the more proficient students introducing items from 'outside' the framework provided on the card, especially concerning experience and qualifications. These comments are then generally picked up and built upon by the interlocutors. It should be pointed out here that the scores obtained on this task are *lower* than those obtained where there is no planning time available. We may speculate that one explanation for this is that learners are using planning time to focus on the content, rather than the accuracy, of the discourse, whereas the raters are more concerned with accuracy than content in the rating procedure. This may be an example of a situation where content competes with accuracy for the limited capacity of the language processor, with content being the winner in the trade-off.

The effect of variable performances by the interlocutor

Variability in the behaviour of interlocutors in assessment situations has in recent years been an area of some concern (see, for example, Ross 1992; Ross and Berwick 1992; Ross 1992; Lazaraton 1992, 1996; Morton et al 1997). The analyses of the discourse in this project also demonstrated that the influence of the interlocutor on the language of the learners could be considerable. For example, the general pace of interlocutor's speech varied across interviews,

as did the clarity with which the speech was presented. Some interlocutors used slow and distinct pronunciation of each word, while others adopted a more naturalistic and fluent speech style. There was much more overlap with the students' utterances in the interactions with some interlocutors than with others, and this can be confusing for the student. In addition, the length of pauses tolerated by the interlocutor in which student can formulate a response varies greatly. All these factors may impact upon the quantity of the student output.

Interlocutors need to consider carefully the complexity of syntax used since this also varies greatly. One approach to reducing this is to specify more precisely the structures which should be adopted in the task rubrics. In addition, interlocutors need to ensure that their selection of lexical items and use of idiomatic expressions remains constant across the range of student assessments undertaken at the same level. It is also essential that interlocutors focus clearly on the task and ensure that their utterances are coherent and relevant. In analysing these data, it appeared that some interlocutors were vague, unfocused and incoherent in a number of cases, with many repairs and reformulations to their speech. This can be confusing and off-putting for learners, particularly those who lack confidence or high levels of proficiency.

It is clear from the findings in this study that the role the interlocutor is expected to adopt should be clearly specified in the task outline. For example, lack of specification of this in some of the tasks meant that certain interlocutors assumed the role of imaginary speakers — such as 'child' or 'workman' — which affected their pronunciation, choice of register etc. Interlocutors also need to take care not to adopt a highly dominant role in dialogues, controlling turn-taking, content, direction and length of dialogues. There was clear evidence in the discourse that this type of behaviour, even though it is probably motivated by the best of intentions, does, in fact, greatly constrain the nature and quantity of the student's output, resulting in a limited sample of language for assessment. The length and scope of dialogues is usually determined by the interlocutor.

On occasion, interlocutors changed the competency being tested through the role they adopted. In some cases, for example, rather than being in the role of seekers of information through requests, students were cast as passive receivers of information which was supplied unsolicited by the interlocutor. While this may result from the interlocutor's awareness of the intended content of dialogue from role cards, it is important that the functional purpose of the compe-

tency is maintained and that the relevant text type is elicited. An example of this is given below. Here the interlocutor proceeds to 'interview' the student over the phone in regard to their prospective suitability for the job, rather than allowing the student to elicit details about the position.

Example 8: [Unstructured task, Certificate III, Competency 5/Task 4]

- I: Good morning. This is Gus Heritage, Coopers Lybrand Consultants.
 S: Good morning, er I would like to know information about the appointment for interview.
 I: Oh, right. Yes, er are you interested in part-time or full-time positions?
 S: Er, full-time position.
 I: Part-time?
 S: Full-time position.
 I: Full-time, full-time. Oh, right, yes, yes. And er do you have any banking or accounting experience?
 S: Er, banking /?/.
 I: Yes, banking experience?
 S: Yeah, yeah.
 I: Yes. Good. Yes. Have you ever worked on telephones before?
 S: Yes, er one ... only one year!
 I: Yes, OK. And er what about your computer experience?
 S: Er /?/, I've finished level 1.
 I: OK. Yes. Well, er, yes, look, if you like I'll send you out a job description, and er you can get your résumé back to me. And then we can contact you if we think you're suitable for an interview.
 S: Yeah.
 I: OK?
 S: Yeah.
 I: Would you like to ask anything else about the job?
 S: Er, maybe not. /?/.
 I: No?
 S: Yeah, thank you.
 I: Well, er, OK. Could I have your name and address, please?
 S: Yeah. (spells name), (says name), (spells name). Er, number seven, number seven, /?/ /(gives suburb)/.
 I: OK. Right. Thank you very much. Goodbye.
 S: Yeah. Goodbye.

Differences in interlocutor approaches to the task may also mean that the task required of some learners is more difficult than that required of other learners, despite the fact that the task is apparently the same. For example, certain interlocutors approached the tasks in a straightforward manner, introducing items that were relevant and central to the task. Others, however, introduced a range of complications from 'outside' that forced the student into a more problematic or unfamiliar area or shifted the focus of the task to another set of skills. It should be pointed out that the introduction of complications is not necessarily to be avoided, but that it should only be provided in cases where it is specifically indicated in the competency statement. What is important is that in any particular task learners receive similar input. This means that where complications are required they should be specified in the task outline. By the same token, where complication or negotiation is specified, it is important that the interlocutor ensures that they are included in the task. For example, in Certificate III, Competency 6, students are required to negotiate a complex or problematic exchange. If, for example, the task requires the student to negotiate a new time for an appointment, and the interlocutor readily agrees, this is not a complex or problematic exchange, but a simple exchange of information. Another student may make a similar suggestion, and the interlocutor refuses, forcing the student to renegotiate or else fail to resolve the situation. This suggests that the approach the interviewer should adopt needs to be specified in the task prompt and that banks of suitable tasks be available for use for assessment purposes.

Inevitably there will be some amount of scaffolding provided in tasks of this type by the interlocutor. Verbal instructions and/or explanations are given by certain interlocutors at the start of the task to reinforce written instructions and this is quite appropriate, but it should be specified that this is always done. However, the amount of assistance provided to the student needs to be limited, and it is important that interlocutors do not provide language models for students they are assessing, either as a lexical item or complete utterances. Differences in interlocutor behaviour are also found where explicit prompts are used (eg telling the student what to do next or what they need to talk about) or where interlocutors pre-empt what they feel the student is trying to communicate by providing the language for them. There are also differences in the overt use of positive reinforcement, eg repeated use of affirmations such as *OK, good, yes, all right, I understand*. Such tokens tend to show support for

the efforts of the students. Thus it is important to provide precise information about the style that should be adopted in the interview. For example, in these data some interlocutors adopt a friendly, relaxed and helpful manner, as in a class pair work activity, rather than a test. Other dialogues are conducted in a more formal manner. It is important to clarify these kinds of issues in the instructions provided to the interlocutor since they may influence student confidence in the task.

Summary and conclusions

In this study, five oral task types at two CSWE ability levels were investigated. Two task characteristics (structure and familiarity) and two task conditions (NS versus NNS and planning time) were manipulated. The following section presents a brief summary of the findings of the study in relation to each of these variables.

Structure

In general, the results for structured versus unstructured tasks had been anticipated — structure makes the task easier, although this is more apparent where the learner is required to obtain information than where the interaction is a negotiated one. We may postulate that the presence of structure in the task reduces the cognitive load placed on the speaker by providing scaffolding upon which to build language. This provides a framework at the global level which in turn allows the learner to pay more attention to the local linguistic content of the response. Of the four tasks in which this variable was manipulated, one involved obtaining information (Certificate III, Competency 5) while the other three required negotiation. The results were most consistent for Certificate III, Competency 5 where the interlocutor role is least specified. In the tasks which require negotiation, the interlocutor role is more precisely specified; we speculate that it may be that this means that the interlocutors, either consciously or unconsciously, provide more structure by their input than is the case in the less negotiable task types. This more structured input would apply regardless of whether the task was a structured one or not, and so may negate the effect of the structure with the result that the effect for structure is less obvious in negotiated exchanges. Thus structure can be built into the task, or provided by the interlocutor. What is important is that there is recognition of the influence it may have.

Familiarity

Due to the problems with the operationalisation of this variable discussed earlier, these results are somewhat speculative. In the Certificate II, Competency 5 task, familiarity appeared to make the task easier. We had expected to find that the more frequently performed (ie familiar) activity would be easier to access and report on, and, although the effect was weak, the more familiar tasks did appear somewhat easier in this competency.

In the Certificate II, Competency 6 task, it appeared that the less familiar task was the easier of the two tasks and in fact this advantage was apparent on all criteria in the raw score analysis. We had expected that obtaining information for oneself in a habitual environment (the English language school) would be easier than obtaining information for a third party in a less familiar environment. There are two points that need to be raised in consideration of this task. First, the task was to some extent compromised by the use of NNS interlocutors for the less familiar task (telephoning the high school), but NS interlocutors for the more familiar task (telephoning the language school). As discussed above, where NS versus NNS was intentionally manipulated as a variable, NNS interlocutors appeared to make the task easier. Thus it may be that the erroneous use of NNS interlocutors in the less familiar task affected the results obtained for the high school task. Second, as alluded to earlier, upon examining the task in more detail, it was found that the less familiar task also included a lot more information about the types of information to be elicited — thus it incorporated more structure into the task than was the case for the more familiar task. This suggests that structure and/or NS versus NNS interlocutors contribute to the outcome. For this reason we may postulate that structure is an important factor in making tasks easier. The issues pertaining to NNS versus NS interlocutors are discussed below.

NS versus NNS interlocutor

Two task conditions were manipulated: NS versus NNS interlocutor, and planning time. The finding for NS versus NNS was not as we had anticipated. We had anticipated that the NS teacher would make the task easier — in fact, this was not the case — the NS interlocutor made the task more difficult in all analyses across both task types. There are a number of reasons for this. First, it may be that the learners are more relaxed in discussion with a NNS interlocutor with the effect that this makes the task easier. Second, raters may compen-

sate for a perceived disadvantage in having a NNS interlocutor. A third alternative is that where the interlocutor is a NNS, the learner tends to produce less (and less complex) language. We may postulate that the NS interlocutor may force the learner to produce more language generally, thus increasing the learner's opportunities to display the full extent of their language competence. However, a substantial range of NNS interviewer proficiencies was used in administering these tasks and careful consideration needs to be given to the effect this variable may have on the learner's performance. Finally, it should also be pointed out that there was a power differential at play here. The NS was a teacher, but the NNS was another learner; thus the social status of the two relationships was not equivalent, and this may have been an intervening variable. In order to identify more precisely the effects of this variable, more detailed investigation is required, and to this end a qualitative analysis of the discourse produced by the tasks is currently being carried out.

Planning time

Finally we turn to planning time. Planning time was not manipulated on its own, but in conjunction with structure in one case, and familiarity of activity in another. It was found that a structured task was easier, and that planning time appeared to increase difficulty in both the structured and the unstructured tasks. A familiar activity was also easier without planning time but, where the task was unfamiliar, planning time had no effect either way. This suggests in general that planning time is not helpful. We may postulate that planning time encourages learners to attempt to introduce more complex ideas, or more complex structures. However, learners' attempts to translate these into linguistic output appear to affect their performance adversely, resulting in the production of less fluent and sometimes less grammatically accurate language. However, it must be noted that these differences are small ones, although we may gain some confidence from the fact that all three approaches to the analysis generally agree over different tasks.

Implications for assessment practice

The outcomes of the research reported in this chapter have a range of implications for assessment practice and for future research. One of the notable findings is that relatively small changes in the characteristics and/or conditions of the task can be shown to influence the scores obtained by learners. Thus it is clearly important to pay considerable attention to very precise specification of

the task at the design stage. In addition, the findings point to the importance of extensive trialling of all oral tasks on a range of learners, and ideally, with a range of interlocutors, before they are used as assessment instruments.

Another message which emerges is that different factors appear to influence different types of tasks to differing degrees. For example, in a task where the learner is required to obtain information, structure appears to be quite influential in its effect on the discourse since it provides a framework for the learner to use. On the other hand, in more negotiated interaction, such as that found in Certificate II, Competency 7 and Certificate III, Competency 6, where questions may be asked and answered both by the learner and the interviewer, structure is not as influential. Here the interlocutor variable is crucial, not only because the interlocutor may provide structure for the learner, but also because this type of task allows much more leeway for the interlocutor to dominate the interaction, to assist or to problematise the task. For this reason, in constructing oral assessments, task designers need to take into account not only the possible variables which can or cannot be incorporated into the task, but also the role of the interlocutor. This role is central in ensuring that learners obtain similar input across similar tasks and so are faced with a similar level of challenge. To this end, training in task design for classroom use needs to include a component aimed at developing teachers' awareness of the ways in which interlocutors can affect tasks both positively and negatively. The more clearly the specifications for assessment task development are spelled out, and the more explicitly the role of the interlocutor and the expectations of his or her input are described, the more reliable the assessment tasks will be.

Finally, although the findings of the project reported here throw some light on the factors which may influence learners' performance on assessment tasks, a good deal of further research remains to be done in order to achieve an understanding of the interactions between task type and task conditions and their impact on an individual's test score. In particular, we need to investigate the way in which specific features of language such as accuracy, fluency and discourse organisation change under differing task conditions and to systematically incorporate this knowledge into the construction and piloting of assessment tasks. Only through continuing research of this kind will it be possible to develop fair and valid assessments of second language performance which reflect the complex range of conditions governing language in use.

Notes

- 1 Post-beginner proficiency is approximately equivalent to Level 1 — on the Australian Second Language Proficiency Ratings (ASLPR), corresponding roughly to Elementary Proficiency to Minimum Survival Proficiency on the Foreign Service Institute (FSI) scale used in the US. Intermediate equates roughly to ASLPR 1+ to 2 (FSI Survival Proficiency to Minimum Social Proficiency).
- 2 For the purposes of the analysis, the 'very easy' category and the 'very difficult' category were collapsed with the 'easy' and 'difficult' categories respectively.

Note: any identifying information has been deleted from all examples, including names, addresses, and phone numbers or part thereof.

References

- Bachman, L 1990. *Fundamental considerations in language testing*. Oxford: Oxford University Press
- Brindley, G 1997. 'Assessment and the language teacher: Trends and transitions'. *The Language Teacher*, 21, 9: 37, 39
- Candlin, C 1987. Toward task-based learning. In C N Candlin and D F Murphy (eds). *Language learning tasks*. Englewood Cliffs, NJ: Prentice Hall
- Crookes, G 1989. 'Planning and interlanguage variation'. *Studies in Second Language Acquisition*, 11: 367–83
- Doughty, C and P Pica 1986. "Information gap" tasks: Do they facilitate second language acquisition?' *TESOL Quarterly*, 20, 2: 305–25
- Douglas, D 1994. 'Quantity and quality in speaking test performance'. *Language Testing*, 11, 2: 125–44
- Ellis, R 1987. 'Interlanguage variability in narrative discourse: Style shifting in the use of the past tense'. *Studies in Second Language Acquisition*, 9: 1–20
- Foster, P 1996. Doing the task better: How planning time influences students' performance. In J Willis and D Willis (eds). *Challenge and change in language teaching*. London: Heinemann

- Foster, P and P Skehan 1996. 'The influence of planning and task type on second language performances'. *Studies in Second Language Acquisition*, 18: 299–323
- Foster, P, A Tonkyn and G Wigglesworth 1998. 'Measuring spoken language: a unit for all reasons'. Paper presented at PACSLRF, March 27–30
- Fulcher, G 1996. 'Testing tasks: Issues in task design and the group oral'. *Language Testing*, 13, 1: 23–52
- Lazaraton, A 1992. The structural organization of a language interview: A conversation analytic perspective. *System*, 1992, 20, 3: 373–386
- Lazaraton, A 1996. 'Interlocutor support in oral proficiency interviews: The case of CASE'. *Language Testing*, 13, 2: 151–72
- Linacre, J M 1990. *FACETS computer program for many faceted Rasch measurement*. Version 2.36. Chicago IL: Mesa Press
- Lumley, T, B Lynch and T McNamara 1994. 'A new approach to standard-setting in language assessment'. *Melbourne Papers in Language Testing*, 3, 2: 19–39
- Messick, S 1989. Validity. In R L Linn (ed). *Educational measurement*. New York: Macmillan
- Morton, J, G Wigglesworth and D Williams 1997. Approaches to the evaluation of interviewer behaviour in oral tests. In G Brindley and G Wigglesworth (eds). *access: Issues in language test design and delivery*. Sydney: National Centre for English Language Teaching and Research, Macquarie University, 175–96
- O'Loughlin, K 1997. Test-taker performance on direct and semi-direct versions of the oral interaction module. In G Brindley and G Wigglesworth (eds). *access: Issues in language test design and delivery*. Sydney: National Centre for English Language Teaching and Research, Macquarie University, 117–46
- Plough, I and S Gass 1993. Interlocutor and task familiarity: effects on interactional structure. In G Crookes and S Gass (eds). *Tasks and language learning: Integrating theory and practice*. Clevedon UK: Multilingual Matters
- Robinson, P 1996. Connecting tasks, cognition and syllabus design. In

- P Robinson (ed). *Task complexity and second language syllabus design: Databased studies and speculations*. Brisbane: University of Queensland Working Papers in Applied Linguistics (Special Issue)
- Robinson, P Forthcoming. Task complexity, cognition and second language syllabus design: A triadic framework for examining task influences on SLA. To appear in P Robinson (ed). *Cognition and second language instruction*. New York: Cambridge University Press
- Robinson, P, S Ting and J Urwin 1995. 'Investigating second language task complexity'. *RELC Journal*, 25, 35–57
- Ross, S 1992. 'Accommodative questions in oral proficiency interviews'. *Language Testing*, 9, 2: 173–86
- Ross, S and R Berwick 1992. 'The discourse of accommodation in oral proficiency interviews'. *Studies in Second Language Acquisition*, 14, 2: 159–76
- Skehan, P 1996. 'A framework for the implementation of task-based instruction'. *Applied Linguistics*, 17, 1: 38–62
- Skehan, P 1998a. Task-based language instruction. In W Grabe (ed). *Annual review of applied linguistics*, 18. New York: Cambridge University Press
- Skehan, P 1998b. *A cognitive approach to language learning*. Oxford: Oxford University Press
- Skehan, P and P Foster 1997. 'Task type and task processing conditions as influences on foreign language performance'. *Language Teaching Research*, 13: 185–211
- Tarone, E 1998. Research on interlanguage variation: Implications for language testing. In L F Bachman and A D Cohen (eds). *Interfaces between second language acquisition and language testing research*. Cambridge: Cambridge University Press
- Varonis, E and S Gass 1985. 'Non-native/non-native conversations: A model for negotiation of meaning'. *Applied Linguistics*, 6: 71–90
- Wigglesworth, G 1997. 'An investigation of planning time and proficiency level on oral test discourse'. *Language Testing*, 14, 1: 101–22

4

Task difficulty and task generalisability in competency-based writing assessment¹

Geoff Brindley

Introduction

This chapter explores the issues of task difficulty and task generalisability in competency-based assessments of second language writing performance in the context of the Certificates in Spoken and Written English (CSWE) used within the Adult Migrant English Program (AMEP) in Australia (see Brindley, Chapter 1, this volume). The study involved an investigation of the writing performances produced by 40 AMEP learners in response to six Certificate III assessment tasks used to assess two writing competencies. The aim of the study was to examine differences in task difficulty within and across competencies and to identify the amount of variability contributed to the competency ratings by tasks and raters. The question of the transferability of common writing skills across tasks was also examined.

Generalisability in performance assessment

Haertel (1993), quoted in Gipps (1994:107) suggests that generalisability in relation to performance assessment can be conceptualised in terms of four levels:

- 1 replicable scoring of a single performance (can we score a single instance of a task in a consistent way?);
- 2 replicability of a specific task (does the performance task have a constant meaning across times and places?);
- 3 generalisability across tasks which are presumed to be assessing the same construct (can we generalise across parallel tasks?); and

- 4 generalisability across heterogeneous task domains (can we generalise across tasks that are not parallel?).

These questions are clearly relevant in the context of the CSWE, since the reliability of the information on learner achievement in the CSWE will clearly be dependent on the quality of the assessment tasks that teachers develop. If outcome reporting is to accurately reflect individual achievement, then the scores yielded by an assessment task developed by a teacher in one location should be able to be compared to those developed by other teachers in other locations. For this reason, as noted previously, variability across raters and tasks should be reduced as far as possible (see Wigglesworth, Chapter 3, this volume). This can be done only by trying to ensure that 1) there is a reasonable level of consistency in the way in which raters classify competencies as achieved or not achieved (see Smith, Chapter 5, this volume) and 2) that assessment tasks aimed at assessing the same competency are parallel in structure and elicit those features of performance that are described in the CSWE performance. In addition, the tasks should, ideally, make similar cognitive demands on learners.

These are very demanding conditions to meet, however, and as yet there is little research evidence from studies of competency-based language assessment to indicate whether these goals are attainable under real-life constraints. The study reported in this chapter therefore aims to throw some light on this question.

Competency and proficiency

Another issue which is of both theoretical and practical significance in the context of the CSWE is the relationship between individual competency achievement and proficiency outcomes. For purposes of reporting learner achievement within the CSWE framework, each competency is reported separately: no attempt is made to translate overall competency achievement into statements of 'general proficiency' describing learners' ability to use the language in non-test situations. As outlined in Chapter 1 of this volume, this is an intentional strategy on the part of the developers of the Certificates who downplay the capacity of CSWE assessments to predict proficiency outcomes (Burrows 1995; Christie and Delaruelle 1997). Nevertheless, it is normally the case that external agencies with a stake in the outcomes of language programs are concerned not so much with classroom task achievement as with transferable skills. It is therefore important to investigate the extent to which the skills demonstrated by

learners in a given CSWE assessment task can be shown to transfer across other tasks, as well as within the same task type. If there were clear relationships across tasks, this might enable links to be made between individual competencies and general proficiency, thus perhaps allowing reporting at an aggregate level. This issue is also of theoretical interest since few attempts have been made in the context of language assessment to identify generalisable dimensions of task and text complexity, either across different types of assessment tasks or within the same task. One way of adding to our knowledge in this area is to investigate the extent to which different competency assessment tasks draw on common components of language ability.

Aims

In the light of the issues discussed above, the present study addresses the following research questions:

- 1 What is the relative difficulty of assessment tasks used to assess the same CSWE writing competency?
- 2 What is the relative contribution of persons, tasks and raters as sources of error variance in competency-based assessments of parallel second language writing tasks?
- 3 How many writing assessment tasks and raters are necessary to achieve acceptable levels of reliability for criterion-referenced decisions concerning achievement of writing competency assessment tasks?
- 4 To what extent can generic components of writing ability be generalised across different writing tasks?

Procedure

Collection and administration of tasks

A set of six CSWE writing assessment tasks was assembled. These consisted of three tasks which had been developed to assess attainment of Competency 10 and three aimed at assessing Competency 12 (Vocational English strand). The tasks were either taken directly from the examples of 'benchmark' tasks included in the CSWE documents or adapted from sample tasks supplied by AMEP provider organisations. The two competencies and the three tasks which accompanied each are described below:

Competency 10: Can write a procedural text**Task 1**

Write a set of instructions on how to use a cassette player.

Task 2

Write a set of instructions on how to enrol in an English course at the Adult Migrant English Service.

Task 3

Write a set of instructions on how to apply for Australian citizenship (leaflet provided).

Competency 12: Can write a report**Task 1**

Write a report about a trade or profession in either your country of origin or Australia.

Task 2

Write a report comparing four hotels (information provided).

Task 3

Write a report about buying a fridge based on a report from *Choice* consumer magazine.

These tasks were then administered to 40 learners enrolled in Certificate III (Vocational English strand) in three states of Australia towards the end of the term at the time when assessment of student achievement would normally take place. Students had spent time studying and practising the features of the two genres concerned but teachers were requested not to prepare them for the specific topics which figured in the assessment tasks.

After doing the tasks, learners were asked to indicate how difficult they had found each task using three simple categories (*easy*, *OK*, *difficult*). Teachers were also asked to comment on their perceptions of the relative difficulty of the tasks.

Rating the scripts

One of the principal aims of the study was to investigate the extent to which the different competency assessment tasks tapped common ability components.

It was therefore necessary to rate the performances according to a set of generic writing criteria in addition to the task-specific CSWE performance criteria. Two separate rating exercises were accordingly undertaken. In the first of these, the scripts were rated by 12 judges from Victoria using a six-point defined rating scale of second language writing ability which had been used previously in other standardised tests of English for adult immigrants (see Delaruelle 1997). This scale was chosen because 1) it had been in use for some time; 2) it had been developed for a similar population in the context of testing for immigration selection and 3) most of the raters (nine out of 12) had had considerable experience in using it. The other three raters were given training in interpreting the criteria and applying the scale before the rating exercise.

Owing to resource constraints it was not possible to have each judge rate each performance. The 240 writing samples were accordingly randomly assigned to the 12 judges in two separate batches so that each script was rated twice according to the four generic criteria on a scale of 0 to 5, with each judge rating 40 scripts.

The features of performance described in the generic scale were as follows:

- Conventions of Presentation (CP)
- Task Fulfilment and Accuracy (TF & A)
- Cohesion and Organisation (C & O)
- Grammatical Control (GC).

In the second rating exercise, a group of 12 experienced CSWE assessors from NSW AMES used the task-specific performance criteria provided in the CSWE to rate the 240 scripts. Once again, scripts were randomised in two batches and each judge rated 40 scripts. The scripts were rated 1 (achieved) or 0 (not achieved) according to each of the following performance criteria as specified in the CSWE:

Competency 12: Can write a report

- Stages report with appropriate beginning, middle and end.
- Writes coherent paragraphs containing factual, clearly organised information.
- Links ideas cohesively as required, eg using conjunctions and reference.
- Uses reference to refer to general categories as required, eg people.

- Uses appropriate vocabulary.
- Uses grammatical structures appropriately, eg simple present and other tenses as required, passive forms.

Methods of analysis

A number of different analytical methods were used. In the first part of the study, many-faceted Rasch analysis, implemented through the computer software package FACETS (Linacre and Wright 1993) was used to examine all of the ratings awarded to each competency and to derive estimates of task difficulty, the relative difficulty of the CSWE performance criteria, as well as measures of rater severity. A generalisability study (G-study) using GENOVA (Crick and Brennan 1984) was then carried out to estimate the magnitude of the variation due to raters and tasks. A series of decision studies (D-studies) was concurrently undertaken to determine the number of tasks and raters necessary to achieve various levels of reliability. Finally, in order to examine the relationship between the generic components of writing ability both within and across the six tasks, raters' scores for all criteria on all competencies were subjected to correlational analysis. Principal factor analysis was then carried out on the resulting correlation matrix to further investigate patterns in the competency ratings.

FACETS and GENOVA runs were carried out on the data derived from both the generic and CSWE ratings. However only the results from the analysis of the CSWE ratings will be reported in addressing Research Questions 1–3, since the generic scheme was used in this study in order to investigate Research Question 4 only and would not normally be used for assessing CSWE performance samples.

Many-faceted Rasch analysis

Background

The Rasch model is one of a family of techniques known as latent trait theory or item response theory (IRT) which have been developed by psychometricians over the last three decades. The basic one-parameter Rasch model states that the probability of a correct response to a given item is a simple logistic function of the difficulty of the item and the ability of the candidate (Pollitt and Hutchinson 1987). This probability is expressed in terms of logits (a shortened form of 'log-odds units' denoting the odds of a candidate getting the item cor-

rect). One of the strengths of the model is that it allows candidate ability and item difficulty to be estimated independently and reported on a common scale, thus avoiding many of the problems associated with sample-dependent classical measurement techniques (Henning 1987).

The many-faceted Rasch model, implemented through the statistical software package FACETS (Linacre and Wright 1993) extends previous Rasch models to include rater characteristics. The FACETS output provides estimates of candidates' ability based on the probability of a candidate obtaining a particular score on a particular task given the ability of the candidate, the difficulty of the item (in the case of writing assessment this might be a rating category such as *conventions of presentation or cohesion*), the harshness of the rater and the effect of any additional facets (Linacre and Wright 1993). The program adjusts candidate ability estimates to take account of raters' tendency to rate either harshly or leniently.

The use of FACETS has proved helpful in the analysis of language test data in a number of ways (see McNamara 1996 for a range of examples). First, it has shifted the focus away from inter-rater agreement to internal consistency. Since the program accepts variability and adjusts ability estimates to take account of rater severity, it is no longer necessary to try to achieve complete agreement between raters. As long as raters are internally consistent, there is no need for them to be excluded from the rating process as has conventionally been the case with raters who did not fit the norm. Second, the FACETS output provides a range of information on rater behaviour which can be used to monitor raters' performance. With the aid of the FACETS output, it is possible to derive estimates of raters' tendency towards severity and leniency and to see how each rater is using the steps on the scale (McNamara and Adams 1991). FACETS also provides item-fit statistics which show the extent to which the information on candidates provided by a given item is consistent with the information from other items, as well as person-fit indices which indicate how well the test is measuring individual candidates. By signalling areas of inconsistent measurement, 'misfit' statistics allow both rater behaviour and rating criteria to be monitored and action taken as necessary. In addition, a technique known as *bias analysis* enables the interactions between different aspects or 'facets' of the rating situation to be examined, eg whether a certain rater is rating candidates of a particular type in a certain way (Wigglesworth 1993). This information can be used to identify inconsistent or biased rating patterns (McNamara 1996).

Results of FACETS analysis

Task difficulty

The task measurement report in Table 19 below shows the FACETS output on the relative difficulty of the CSWE tasks. It can be seen that the three tasks assessing Competency 10 and those assessing Competency 12 were not of equal difficulty. However, the differences amount to only .57 of a logit in the case of Competency 10 and slightly less where Competency 12 is concerned. The step separation of 1.52 is also quite small, suggesting that there are not large differences in difficulty between the tasks.

Tasks 1 and 2 in Competency 10 show a standardised infit mean-square statistic of 2, indicating inconsistencies in rating patterns. Competency 12, Task 1 is signalled as significantly overfitting, indicating a lack of variation in the ratings.

Competency 12, Task 1 which required learners to write a report on a trade or profession in their country was the most difficult task while the easiest was Competency 12, Task 2, a comparison of four hotels.

Table 19: CSWE Writing — Competencies 10 and 12: Task Measurement Report

Obsvd Score	Obsvd Count	Obsvd Avrge	Fair Avrge	Measure Logit	Model Error	Infit MnSq	Std	Outfit MnSq	Std	N	Task
190	312	0.6	0.5	-0.01	0.14	1.1	2	1.2	1	1	Competency 10 Task 1
182	316	0.6	0.4	0.20	0.14	1.1	2	1.2	1	2	Competency 10 Task 2
196	304	0.6	0.6	-0.37	0.14	1.0	0	1.9	4	3	Competency 10 Task 3
229	462	0.5	0.4	0.27	0.11	0.9	-2	0.9	-1	4	Competency 12 Task 1
245	462	0.5	0.6	-0.21	0.11	1.0	0	0.9	-1	5	Competency 12 Task 2
248	464	0.5	0.5	0.12	0.11	0.9	-1	0.8	-2	6	Competency 12 Task 3
215.0	386.7	0.6	0.5	0.00	0.12	1.0	0.0	1.2	0.4	Mean (Count: 6)	
26.6	76.1	0.1	0.1	0.23	0.01	0.1	1.8	0.4	2.4	S.D.	

RMSE 0.12 Adj S.D. 0.19 Separation 1.52 Reliability 0.70
 Fixed (all same) chi-square: 19.9 d.f.: 5 significance: .00
 Random (normal) chi-square: 4.9 d.f.: 4 significance: .29

Rater severity

The rater measurement report in Table 20 below shows the FACETS estimates of the relative severity of the CSWE raters. The table shows that the CSWE

raters spanned a range of nearly three logits, -1.48 to 1.43. Two (Raters 8 and 10) were flagged by the program as misfitting (standardised infit of 2 or more), indicating inconsistent rating patterns, while two demonstrated significant negative fit statistics of -2, suggesting little variation in the use of scale points. On this point, Engelhard (1992:178) comments that ‘raters with muted rating patterns tend to score holistically rather than analytically’. The separation index, which indicates the extent to which raters are reliably different in their levels of severity, was 4.26, indicating substantial differences in severity.

Table 20: CSWE Writing — Competencies 10 and 12: Rater Measurement Report

Obsvd Score	Obsvd Count	Obsvd Avrge	Fair Avrge	Measure Logit	Model Error	Infit MnSq	Std	Outfit MnSq	Std	N	Rater
115	196	0.6	0.5	-0.02	0.16	1.0	0	1.3	2	1	
128	186	0.7	0.7	-0.86	0.17	1.0	0	1.0	0	2	
108	188	0.6	0.5	-0.06	0.17	1.1	1	1.2		3	
110	198	0.6	0.5	0.10	0.16	0.9	-2	0.8	-2	4	
120	200	0.6	0.5	-0.18	0.17	1.0	0	1.0	0	5	
123	190	0.6	0.6	-0.45	0.17	0.9	-1	0.8	-1	6	
155	198	0.8	0.8	-1.48	0.19	0.9	0	1.1	0	7	
92	206	0.4	0.4	0.29	0.16	1.2	3	1.3	2	8	
74	188	0.4	0.2	1.14	0.18	0.9	-1	0.8	-1	9	
112	188	0.6	0.6	-0.23	0.17	1.2	2	2.4	6	10	
99	186	0.5	0.4	0.32	0.17	0.8	-2	0.7	-2	11	
54	196	0.3	0.2	1.43	0.18	1.0	0	0.9	0	12	
107.5	193.3	0.6	0.5	0.00	0.17	1.0	-0.1	1.1	0.4	Mean (Count: 12)	
24.9	6.2	0.1	0.2	0.75	0.01	0.1	1.6	0.4	2.4	S.D.	

RMSE 0.17 Adj S.D. 0.73 Separation 4.26 Reliability 0.95
 Fixed (all same) chi-square: 205.9 d.f.: 11 significance: .00
 Random (normal) chi-square: 10.9 d.f.: 10 significance: .36

Relative difficulty of performance criteria

The item measurement report in Table 21 below shows the FACETS difficulty estimates for each of the performance criteria evaluated in the CSWE assessment tasks. The analysis shows that the rating category of *sequence (clearly signals sequences linguistically, symbolically or numerically as required)* was flagged by FACETS as misfitting, with a standardised positive infit of two. This

may reflect some differences in raters' interpretation of what constitutes a 'clear signal' of sequence. Two other performance criteria, *CohPar* (*writes coherent paragraphs containing factual, clearly organised information*) and *Reference* (*uses reference to refer to general categories as required*) were signalled as significantly overfitting, signalling a lack of variation in the ratings. One possible interpretation of these overfitting categories is that they are not independent from the others (Linacre and Wright 1993:65; McNamara 1996:222ff). Raters, in other words, may be conflating the assessment categories. It could be argued here, for example, that the ability to write coherent paragraphs which refers to 'clearly organised information' may overlap in the minds of raters with the *staging* since criterion staging requires clear organisation; the evidence for overlap is a little stronger in the case of the *reference* criterion where the *cohesive links* criterion explicitly gives reference as an example of ability to link ideas cohesively, creating potential duplication of rating criteria. However, further empirical investigation of rating processes *in situ* would be necessary to establish to what extent this interpretation could be sustained.

The most difficult rating category on which to gain a high score was *grammatical structure* with a logit value of .77, while the easiest was *sequence* (-1.14 logits). The difference of nearly two logits in the difficulty of the rating categories, combined with the step separation of 4.4 (reliability .95), suggest that the performance criteria differ substantially in difficulty.

Table 21: CSWE Writing — Competencies 10 and 12: Item Measurement Report

Obsvd Score	Obsvd Count	Obsvd Avrge	Fair Avrge	Measure Logit	Model Error	Infit MnSq	Std	Outfit MnSq	Std	N	Item
247	464	0.5	0.5	0.19	0.11	1.1	1	1.1	1	1	Staging
102	232	0.4	0.3	0.64	0.15	0.9	-2	0.8	-2	2	CohPar
117	231	0.5	0.4	0.26	0.15	0.9	-1	0.8	-1	3	CohesLinks
134	232	0.6	0.5	-0.11	0.15	0.9	-2	0.8	-1	4	Reference
314	464	0.7	0.6	-0.61	0.11	1.0	0	1.1	1	5	Vocab
197	464	0.4	0.3	0.77	0.11	1.0	0	1.0	0	6	Gramm
179	233	0.8	0.8	-1.14	0.17	1.2	2	2.0	3	7	Sequence
184.3	331.4	0.6	0.5	0.00	0.14	1.0	-0.3	1.1	0.1		Mean (Count: 7)
70.5	114.8	0.1	0.1	0.63	0.03	0.1	1.5	0.4	2.0		S.D.

RMSE 0.14 Adj S.D. 0.62 Separation 4.40 Reliability 0.95
 Fixed (all same) chi-square: 143.7 d.f.: 6 significance: .00
 Random (normal) chi-square: 6.0 d.f.: 5 significance: .31

Generalisability theory (G-theory) analysis

Background

The next part of the analysis consisted of an examination of the relative contribution of variation in tasks and rater judgements to variation in the competency ratings. This was carried out using generalisability theory (G-theory), which provides a way in which researchers can evaluate the effects of multiple sources of variance, known as *facets*, on test scores (Brennan 1992). Facets may include tasks, raters, items or scoring occasions. The estimated variance components associated with the facets and object of measurement (here *persons*) which are derived from a generalisability study (G-study) indicate the amount of error in generalising from a person's score from a single rater on a single task to his or her so-called *universe score*, or average score over all possible raters and tasks in the universe. The G-study variance components are subsequently used as input to estimate decision study (D-study) variance components. These provide information about the effects of different combinations of facets on the dependability of the scoring system in the form of an index of *generalisability* (the G-coefficient) for relative (norm-referenced) decisions and *dependability* (the *phi* coefficient) for absolute (criterion-referenced) decisions. In this way it is possible to estimate how many tasks, raters or items would be necessary to obtain given levels of dependability. G-theory, in combination with many-faceted Rasch analysis, has been used for these purposes in analyses of language test data by Bachman et al (1996) and McNamara and Lynch (1997) and Lynch and McNamara (1998).

Study design

In this study, the universe of admissible observations consisted of three facets: raters, tasks and items. Raters and tasks were considered to be random in the analysis since they were sampled from larger universes and the intent of the study was to make generalisations about larger domains. The performance criteria (items) from each competency were treated as a fixed effect, however, since they do not vary according to the assessment task used. When a facet is fixed in the D-study, the variance components for the interaction between items and persons is treated by GENOVA as a component of the universe-score variance estimate and is excluded from further estimation procedures (Lane and Sabers 1989:200). For this reason, item-related interactions will not be reported in the results of the GENOVA analysis.

Analysis

The analyses were carried out using GENOVA for the Macintosh computer, the statistical software program through which generalisability theory is implemented (Crick and Brennan 1984). The focus here was on absolute decisions since one of the aims of the study was to ascertain whether or not candidates met a particular performance standard (that is, whether or not they would be accredited with achievement of a particular competency) and it was therefore the dependability coefficient (*phi*) which was of interest. Results of the analysis for each of the two writing competencies are presented below. The full G-study variance components are only reported for the combination of a single task and a single rater since this reflects the application of the CSWE rating system under normal conditions. The summaries of the D-studies, with dependability coefficients for different combinations of tasks and raters are provided in Tables 23 and 25.

Results

Competency 10

The results of D-study 1 (1 rater x 1 task x 4 items) are shown in Table 22. Using a single task and a single rater, the variance component for persons was only 21.27%. The person-by-task component accounted for 24.55% of total variance, suggesting that candidates were rank-ordered differently by the tasks, while the largest component is the residual error of 54.18 which includes all of the person-by-task by-rater interactions and other unexplained sources of error variance. The fact that candidate ability accounts for such a small proportion of the variance clearly indicates that dependable measures cannot be obtained under these conditions.

The person-by-rater and rater-by-task interactions are both zero, which would suggest that raters rank order persons and tasks similarly. However, this is in contrast to the large differences in rater severity revealed by the FACETS output. This point will be taken up further on in the discussion of results.

Table 22: Competency 10: Variance Components for D-study 1
(1 rater x 1 task x 4 items)

Effect	Variance component	Standard error	% variance
Persons (P)	0.0316	0.01298	21.27
Tasks (T)	0.0000	0.00061	0.00
Raters (R)	0.0000	0.00017	0.00
PT	0.03482	0.01308	24.55
PR	0.0000	0.00572	0.00
TR	0.0000	0.00031	0.00
PTR	0.07684	0.01215	54.18

Competency 10: D-studies

Table 23 summarises the sixteen D-studies conducted on Competency 10. It sets out the effects on dependability of using differing combinations of tasks and raters. It can be seen that the dependability coefficient for criterion-referenced decisions is very low (.213) when a single task is assessed by a single rater. With six tasks and one rater it rises to .618. However, six tasks and two raters would be necessary to obtain a dependability of .71, a level which would be considered minimal for low-stakes decisions.

Table 23: Competency 10: Dependability (*phi*) coefficients for D-studies

Tasks	Raters	Items	<i>Phi</i>	Tasks	Raters	Items	<i>Phi</i>
1	1	4	0.21267	1	3	4	0.33292
2	1	4	0.35075	2	3	4	0.49954
3	1	4	0.44763	3	3	4	0.59956
4	1	4	0.51934	4	3	4	0.66626
5	1	4	0.57458	5	3	4	0.71391
6	1	4	0.61843	6	3	4	0.74965
1	2	4	0.29169	1	4	4	0.35824
2	2	4	0.45164	2	4	4	0.52751
3	2	4	0.5526	3	4	4	0.62612
4	2	4	0.62225	4	4	4	0.69068
5	2	4	0.67310	5	4	4	0.73623
6	2	4	0.71189	6	4	4	0.77008

Competency 12

Table 24 below shows the variance components for D-study 1 (one task and one rater). As with Competency 10, the GENOVA analysis shows a large residual variance component. Although the person by task variance component is considerably smaller than in the case of Competency 10, accounting for 10.78% of the total variance, it nevertheless still indicates a difference in the rank ordering of person scores for the tasks.

Table 24: Competency 12: Variance components for D-study I
(1 rater x 1 task x 6 items)

Effect	Variance component	Standard error	% variance
Persons (P)	0.05895	0.01726	41.42
Tasks (T)	0.00000	0.00061	0.00
Raters (R)	0.00102	0.00134	0.70
PT	0.01534	0.00911	10.78
PR	0.00228	0.00628	1.60
TR	0.00000	0.00040	0.00
PTR	0.06473	0.01023	45.48

Competency 12: D-studies

The summary of the D-studies on Competency 12 shown in Table 25, however, presents a somewhat different picture to the analyses of Competency 10. Although dependability with one task and one rater is still quite low at .414, with two tasks and two raters it rises to .698, sufficient for low-stakes decisions. Four tasks and one rater would achieve a dependability of .717.

Table 25: Results of D-studies for Competency 12

Tasks	Raters	Items	<i>Phi</i>	Tasks	Raters	Items	<i>Phi</i>
1	1	6	0.41419	4	2	6	0.81278
2	1	6	0.57630	5	2	6	0.84041
3	1	6	0.66276	6	2	6	0.85991
4	1	6	0.71651	1	3	6	0.60794
5	1	6	0.75316	2	3	6	0.75087
6	1	6	0.77974	3	3	6	0.81471
1	2	6	0.54429	4	3	6	0.85088
2	2	6	0.69801	5	3	6	0.87417
3	2	6	0.77054	6	3	6	0.89042

Relationship of components of writing competence

Analysis

The aim of the final part of the study was to investigate the extent to which the different competency assessment tasks drew on common underlying writing abilities described in the generic scale (*conventions of presentation, task fulfilment and appropriacy, cohesion and organisation and grammatical control*).

To this end, Pearson correlations were computed between the scores on the four assessment categories within each task and across tasks. The results are shown in Tables 26a and 26b. For ease of reference, the separate correlation matrices are shown for Competency 10 and Competency 12.

Although there are mid-range correlations for some of the components across tasks, particularly in the case of Competency 12, it can be seen that the correlations between the within-task categories are uniformly higher. Of the total correlations between competence components, 83% are above .7 and 44% above .8.

Table 26a: CSWE Writing competency 10: Correlation matrix

	CP101	TFA101	CO101	GC101	CP102	TFA102	CO102	GC102	CP103	TFA103	CO103	GC103
CP101	1.000											
TFA101	0.731**	1.000										
CO101	0.705**	0.807**	1.000									
GC101	0.773**	0.766**	0.813**	1.000								
CP102	0.394*	0.458*	0.380*	0.434*	1.000							
TFA102	0.497**	0.542**	0.445**	0.475**	0.691**	1.000						
CO102	0.369*	0.577**	0.405*	0.377*	0.711**	0.859**	1.000					
GC102	0.318*	0.284	0.140	0.313	0.671**	0.595**	0.677**	1.000				
CP103	0.318*	0.388*	0.307	0.299	0.470**	0.408*	0.336*	0.481**	1.000			
TFA103	0.269	0.421**	0.365*	0.242	0.380**	0.398*	0.364*	0.429*	0.747**	1.000		
CO103	0.351*	0.482**	0.409*	0.286	0.287	0.297	0.213	0.261	0.766**	0.857**	1.000	
GC103	0.396*	0.476**	0.464**	0.444*	0.379*	0.319*	0.264	0.401*	0.822**	0.769	0.788	1.000

**p = < .01

*p = < .05

Key:

- CP: Conventions of presentation
 TFA: Task fulfilment and appropriacy
 CO: Cohesion and organisation
 GC: Grammatical control
 101: Competency 10, Task 1
 102: Competency 10, Task 2
 103: Competency 10, Task 3

Table 26b: CSWE Writing competency 12: Correlation matrix

	CP121	TFA121	CO121	GC121	CP122	TFA122	CO122	GC122	CP123	TFA123	CO123	GC123
CP121	1.000											
TFA121	0.793**	1.000										
CO121	0.860**	0.877**	1.000									
GC121	0.872**	0.836**	0.836**	1.000								
CP122	0.443**	0.434*	0.396*	0.546**	1.000							
TFA122	0.503**	0.535**	0.502**	0.532**	0.663**	1.000						
CO122	0.577**	0.493**	0.515**	0.596**	0.722**	0.832**	1.000					
GC122	0.558**	0.481**	0.504**	0.609**	0.499*	0.706**	0.817**	1.000				
CP123	0.549**	0.493**	0.523**	0.542**	0.555**	0.398*	0.467**	0.286	1.000			
TFA123	0.576**	0.519**	0.540**	0.505**	0.448**	0.406**	0.481**	0.295	0.885**	1.000		
CO123	0.509**	0.470**	0.455**	0.447*	0.438**	0.392**	0.445**	0.291	0.855**	0.932**	1.000	
GC123	0.501**	0.458*	0.497**	0.431**	0.391*	0.368*	0.442**	0.317*	0.795**	0.857**	0.791**	1.000

**p = < .01

*p = < .05

Key:

- CP: Conventions of presentation
 TFA: Task fulfilment and appropriacy
 CO: Cohesion and organisation
 GC: Grammatical control
 121: Competency 12, Task 1
 122: Competency 12, Task 2
 123: Competency 12, Task 3

To further investigate the scoring patterns across the different competency assessment tasks and to examine the way in which the rating criteria were functioning, a Principal Factor Analysis was conducted on the complete correlation matrix for all tasks, using orthogonal (Varimax) rotation. This yielded a six-factor solution explaining 81.75% of the total variance (Table 27) in which the within-task competence components from each task loaded heavily together, with each task appearing to constitute a separate factor (the six factors are indicated in bold). This would seem to suggest either 1) that raters are not differentiating clearly between the rating criteria and are reacting to the task as a whole or 2) that there is relatively limited transfer of skills across tasks. The relative merits of each of these interpretations will be considered in the discussion section below.

Limitations of the study

Before considering the significance of the results of these analyses, it is important to point out a number of limitations of the study design and the analytical tools used.

In the first place, it could be argued that some of the criteria used in the generic scale did not specifically cover the features of the task at hand and were therefore difficult to apply to some of the text types being assessed. This may have led to some irregularities or inconsistencies in the application of the rating criteria, although the FACETS item misfit statistics signalled only seven unexpected responses for the generic rating scheme (as opposed to 25 for the dichotomous CSWE rating scheme). Although these potential problems were foreseen and a set of ground rules formulated prior to the rating session concerning the interpretation of the criteria, there is always the possibility that the adaptation of a generic scale to some texts may have proved difficult for some raters. The application of a criterion such as ‘cohesion and organisation’, for example, to a procedural text which simply calls for candidates to list a series of steps in a given order requires careful consideration and was the subject of some discussion amongst raters prior to and after the first rating session.

Second, the extent to which generalisation is possible across parallel tasks is dependent on the type and amount of prior instruction received by learners. While considerable efforts were made to control the conditions under which learners attempted the tasks, some of the teachers involved stated that they had prepared students for the content of the tasks in advance, although they had

Table 27: Competencies 10 and 12: Principal factor analysis

	1	2	3	4	5	6
CO122	0.918	0.016	0.242	0.129	0.089	0.243
TFA122	0.760	0.092	0.162	0.185	0.091	0.280
GC122	0.721	-0.032	0.059	0.069	0.286	0.371
CP122	0.606	0.371	0.255	0.147	0.102	0.219
CO102	-0.074	0.898	0.080	0.114	0.265	-0.060
TFA102	0.032	0.795	0.048	0.173	0.334	-0.044
GC102	0.155	0.729	0.254	0.200	0.035	0.236
CP102	0.189	0.722	0.094	0.196	0.235	0.006
TFA123	0.166	0.104	0.910	0.175	0.085	0.269
CO123	0.192	0.083	0.873	0.167	0.054	0.197
CP123	0.217	0.185	0.833	0.104	-0.009	0.272
GC123	0.141	0.136	0.773	0.175	0.168	0.251
CO103	0.091	0.052	0.194	0.866	0.226	0.090
TFA103	0.094	0.269	0.060	0.842	0.111	0.231
GC103	0.149	0.125	0.137	0.803	0.294	0.134
CP103	0.190	0.278	0.200	0.793	0.105	-0.002
CO101	0.068	0.142	0.005	0.235	0.862	0.003
GC101	0.216	0.221	0.051	0.096	0.852	0.047
TFA101	0.033	0.314	0.056	0.297	0.789	-0.155
CP101	0.164	0.215	0.156	0.120	0.772	0.067
CO121	0.239	0.012	0.263	0.173	-0.105	0.872
CP121	0.293	0.063	0.302	0.066	0.061	0.812
TFA121	0.269	-0.002	0.260	0.135	-0.064	0.810
GC121	0.364	0.046	0.227	0.095	0.090	0.808
Eigenvalue	9.427	4.236	1.906	1.707	1.384	0.961
% Variance	12.531	12.902	14.691	13.582	13.589	14.459

Total % of variance explained: 81.754

Key:

- CO: Cohesion and organisation
- CP: Conventions of presentation
- TFA: Task fulfilment and appropriacy
- GC: Grammatical control
- 101: Competency 10, Task 1
- 102: Competency 10, Task 2
- 103: Competency 10, Task 3
- 121: Competency 12, Task 1
- 122: Competency 12, Task 2
- 123: Competency 12, Task 3

been specifically requested not to do so. Differences in task difficulty may therefore have reflected to some extent these differing degrees of preparation.

Third, it should be acknowledged that the numbers of learners (40) and tasks (6) included in the analysis are relatively small and the logit estimates for the different facets may therefore not be stable. Because of the data requirements for FACETS, it was the original intention of this study to gather a complete set of ratings on the six tasks for a total of 120 learners. Unfortunately this turned out to be impossible due to a high student absence rate on the one hand and the discontinuation of government funding for the programs which provided the bulk of the assessment data on the other. As a result there were missing scores for over 50 candidates, resulting in a considerably smaller data set than had originally been envisaged. In this regard, a sample of at least 100 subjects is frequently suggested as a minimum for the use of the Rasch model (McNamara 1996), although some studies (eg Lumley et al 1994) have employed samples as small as 20. Wright and Tennant (1996:468) argue that 'with a reasonably targeted sample of 50 persons, there is a 99% confidence that the estimated item difficulty is within ± 1 logit of its stable value, especially when persons take 10 or more items'.

Discussion

These limitations notwithstanding, a number of tentative interpretations can be advanced in relation to the research questions under investigation.

Task difficulty

The first research question concerned whether or not the tasks which aimed to assess the same competency were of equivalent difficulty.

That the tasks were of differential difficulty was seen in the large person-by-task variance component shown by the GENOVA results and the difference in logit values seen in the FACETS analysis. The difference in the logit values, however, was not a substantial one. This can be seen by calculating the probabilities of candidate success on a task of a given difficulty taking into account rater severity, using the logit-to-probability conversion table provided by Wright and Linacre (1991). In the case of Competency 10 — where the tasks showed the largest differences — candidates would have about a 14% better chance of successfully achieving Competency 10 if they were given Task 3 rather than Task 2. In assessing achievement of Competency 12, if a teacher

administered Task 1 instead of Task 2 the candidate's chances of success would be reduced by about 12%. Since the CSWE assessments at this point do not involve particularly high stakes, the consequences of this degree of difference in task difficulty may not be serious, although it could be argued that some Certificate III learners could be disadvantaged if they were denied certification on the basis of an assessment by a more severe rater.

It was not the aim of this part of the study to identify specific reasons for the differences in task difficulty. On closer inspection of the characteristics of the individual tasks a number of possible explanations could be tentatively advanced which are outlined briefly below. These would, however, need to be explored through further research into the rating process itself.

As mentioned previously, differential amounts of preparation may have had an influence on performance, albeit an unintended one, since efforts were made to control this variable by issuing guidelines regarding the type of preparation to be given. Another key factor is the degree of involvement of skills other than those being assessed. In this regard, it is perhaps significant that Competency 12, Task 3 (writing a report on the relative merits of different brands of refrigerators) which required learners to read a considerable amount of lengthy and quite dense input material, was universally identified by both teachers and learners as the most difficult and demanding task. Interestingly, however, the FACETS analysis did not support this perception. On the one hand, this may indicate that the task was more manageable for learners than they themselves and their teachers had estimated. Alternatively, it is possible that teachers may have compensated for the perceived difficulty of the task and awarded higher ratings (see Hamp-Lyons and Mathias 1994 who advance this explanation for large discrepancies between teachers' estimations of task difficulty and actual student performance as evidenced in item statistics). Another possible explanation is that learners were successfully able to incorporate material from the input text and therefore gained higher scores than they would have without the support material. In any case, the differences between the tasks in terms of the support materials provided indicate the need for clear guidelines regarding the type and amount of reading input to be given for writing assessment tasks, if the tasks are intended to be parallel.

Rater severity

Although differences in task difficulty were not large, the FACETS output showed major differences in rater severity in both rating schemes. This result

was in contrast to the zero variance component for rater-by-person and rater-by-task interactions revealed by the GENOVA analysis. Other studies using G-theory and FACETS to investigate language test data have found a similar phenomenon (Bachman et al 1995a; Lynch and McNamara 1998). Bachman et al (1995a), for example, in an investigation of the speaking subtest of a university level Spanish proficiency test, found a difference of more than four logits in ratings of oral tasks from the FACETS analysis but a zero variance component for raters in their G-study. They explain this apparent discrepancy by noting that FACETS is more sensitive to individual differences than GENOVA and suggest that in this case, a range of individual rater severity is unlikely to produce an overall effect on test scores. In a similar type of study, Lynch and McNamara (1998) investigated the performances of 83 adult ESL learners in a test of speaking skills. They found a difference of almost 2.5 logits in rater severity which was reflected in a substantial variance component (8.5% of total variance) for raters. On the other hand, there was a quite small rater-by-person interaction when compared with extensive rater-person bias revealed through a FACETS bias analysis. The researchers explain these differing results thus:

One way of reconciling these differences is to recognize that the GENOVA and FACETS analyses operate with differing levels of detail. Using the microscope as an analogy, FACETS turns up the magnification quite high and reveals every blemish on the measurement surface. GENOVA, on the other hand, sets the magnification lower and tends to show us only the net effect of the blemishes at the aggregated level. This is not to say that 'turning up the magnification' is the same as increasing the accuracy. It merely suggests that there is a different level of focus (individuals versus groups). In other words, while the FACETS analysis identified a great number of specific person-by-rater and rater-by-item combinations that were considered to be 'biased', the GENOVA analysis indicated that the effect of these combinations were likely to wash out across all the ratings. (1998:176)

These comments notwithstanding, the FACETS analysis would nevertheless seem to suggest that an individual candidate's chances of being awarded the competency in question (and hence the CSWE) would be heavily affected by the severity of the rater who rates their performance. This can be seen by calculating the probabilities of candidate success, using the Wright and Linacre (1991) conversion table referred to above. Here the difference in severity

between the CSWE raters of almost three logits means that a candidate's chances of being awarded Competency 10 or 12 would be reduced by approximately 45% if his or her script were judged by the most severe rater instead of the most lenient. This highlights the importance of using more than one rater if the results of the assessment are to be used to make important decisions.

Difficulty of performance criteria

Similar to other studies which have used Rasch analysis to investigate the use of scale categories by raters (cf Brown 1995; McNamara 1996), this study revealed a hierarchy of difficulty among the performance criteria used to assess writing performance. Grammar was the most harshly rated category, a finding which is also consistent with other language testing studies (McNamara 1996:223). Some of the rating criteria elicited 'muted' rating patterns which suggests that raters may not be distinguishing between some of the rating categories. This could be partially due to halo effect whereby the rater carries out an initial overall impression rating and then transfers this level of performance to all of the criteria. Alternatively, some of the performance criteria may be overlapping, with some performance features subsuming aspects of others. However, without evidence from raters on the ways in which they interpret and use the criteria, it is difficult to reach any definitive conclusions. In the meantime, analysis of the type reported here can identify potentially problematic criteria (through, for example, the Rasch misfit statistics). These can then be investigated further through discussions with raters and modified as necessary.

Dependability

The second research question concerned the relative contribution of persons, tasks and raters as sources of error variance. If tasks are intended to be replicable, it is clearly desirable that task sampling variability be kept to a minimum in order to enable dependable measurement across tasks, occasions and raters.

Under normal conditions of administration of the CSWE, that is, the administration of one task by one rater, the generalisability study revealed low levels of dependability for both competencies. However, the D-studies indicate that the addition of tasks increases the dependability coefficient, in some cases markedly so. Here it is worth noting that there were considerable differences between Competency 10 and Competency 12 in terms of rater consistency, with the former showing a much lower level of dependability. This finding was corroborated by the FACETS analysis which identified two of the three tasks

used to assess Competency 10 as misfitting, suggesting that these tasks are eliciting inconsistent rating patterns. With Competency 10, four tasks and four raters would be necessary to obtain a dependability coefficient of .71, and even with six tasks, *phi* still only reaches .61 if only one rater is used. On the other hand, as far as Competency 12 is concerned, a *phi* coefficient of .7 can be obtained with two tasks and two raters, a situation which could conceivably be envisaged under normal operational conditions. This suggests that it is possible with some competencies to obtain acceptable levels of dependability for curriculum-related decisions.

Generalisability of writing ability components across different tasks

The third research question concerned the extent to which the underlying components of writing exemplified in the generic rating scheme were generalisable across different competency assessment tasks.

On the face of it, the evidence from this study would seem to suggest that the competence components do not generalise across the two writing competencies under investigation. This is evidenced by high and uniform correlations between the ratings for the components within each of the six tasks, contrasted with low to moderate across-task correlations. In addition, the results of the principal factor analysis show the competence components in each of the six tasks loading heavily together on a single factor. Contrary to the findings of other studies which have found reasonably strong relationships between competence components across different kinds of writing tasks (eg Pollitt and Hutchinson 1987), the tasks here appear to be behaving almost independently.

However, it would be premature to conclude that this finding constitutes evidence for the task-specificity of language skills. The fact that the generic writing skills seem to cluster together within each task is likely to be indicative of a strong halo effect, as discussed above, rather than of lack of skills transfer. It may be, in other words, that raters are strongly influenced by their initial overall impression of the quality of the text and do not subsequently pay close attention to each individual feature of performance. Once they have decided that the performance is an overall '3', all of the criteria — perhaps unconsciously — are then brought into line with this rating. Further evidence in support of this explanation is provided by the FACETS item measurement report for the generic rating scheme shown in Table 28 below where the four rating

categories showed a separation index of only 1.52 with a reliability of .7. This suggests that the categories are not reliably different.

Table 28: Rating categories: Generic rating scheme — Item measurement report

Obsvd Score	Obsvd Count	Obsvd Avrge	Fair Avrge	Measure Logit	Model Error	Infit MnSq	Std	Outfit MnSq	Std	N	Item
1517	480	3.2	3.0	0.06	0.06	1.0	0	1.0	0	1	TF&A
1535	480	3.2	3.0	-0.01	0.06	1.0	0	1.0	0	2	CP
1577	480	3.3	3.1	-0.18	0.06	1.1	1	1.1	1	3	C&O
1498	480	3.1	3.0	0.13	0.06	0.9	-1	0.9	-1	4	GC
1531.8	480.0	3.2	3.0	0.00	0.06	1.0	-0.0	1.0	-0.0	Mean (Count: 4)	
29.2	0.0	0.1	0.1	0.11	0.00	0.1	1.2	0.1	1.2	S.D.	

RMSE 0.06 Adj S.D. 0.09 Separation 1.52 Reliability 0.70

Fixed (all same) chi-square: 13.2 d.f.: 3 significance: .00

Random (normal) chi-square: 3.0 d.f.: 2 significance: .22

Interestingly, McNamara and Lynch (1997:209) who conducted a generalisability study of test scores derived from the same rating scale used in the present study, report a similar phenomenon:

The results indicate the presence of a fairly strong halo effect, whereby as raters judge different aspects of the same piece, each rating influences the likelihood of the next.

This halo effect may be a result of the conditions under which ratings are typically carried out. In this connection, Vaughan (1991:121) points out that when samples of student writing are read quickly, one script may blur into another and ratings are made by comparing the sample at hand with the previous one, rather than through reference to a set of rating guidelines. In a similar vein, Henning (1988) argues that juggling multiple assessment categories simultaneously places a heavy cognitive load on readers, contributing to a halo effect. To counter this effect, Hamp-Lyons and Henning (1991:364) suggest that 'scores are more likely to demonstrate separability if each were assigned on a separate scoring occasion', a suggestion also advanced by McNamara and Lynch (op cit). As Hamp-Lyons and Henning acknowledge, however, such a procedure would be impractical since writing tasks almost always have to be judged on a single occasion, although it would be of undoubted interest as a research exercise.

Summary and conclusions

The first research question concerned whether or not tasks used to assess the same writing competency were of equivalent difficulty. Here, although Rasch and G-theory analyses revealed the tasks to be differentially difficult, these differences were not large. By contrast, considerable differences in rater severity were revealed by the Rasch analysis. The individual performance criteria within the two competencies were also revealed to be substantially different in difficulty. Although this situation is fairly typical where performance assessments are concerned (McNamara 1996:234), these differences would need to be controlled by using multiple rating procedures or ratings adjusted for severity in cases where the results of the assessments were to be used for high-stakes decisions.

The second and third research questions concerned the relative contribution of persons, tasks and raters to variation in assessment scores and the number of raters and tasks that would be necessary to achieve acceptable levels of dependability in competency-based assessments. Here, the G-theory analyses revealed substantial measurement error and low levels of dependability under normal assessment conditions, that is, when a single rater assesses a competency using a single task. However, the results of the D-studies suggested that, in the case of one of the two writing competencies under investigation, it would be possible to obtain acceptable levels of dependability for low-stakes decisions using two tasks and two raters, a situation which could be envisaged under operational conditions.

The fourth research question addressed the degree to which the different competency assessment tasks tapped common elements of writing ability. Ratings of the same 40 writing scripts were obtained using a scale which incorporated four 'generic' elements of writing ability.

Correlational and factor analyses of ratings of tasks suggested relatively weak relationships between components of writing ability across tasks. High within-task correlations and a small step separation index revealed by the Rasch analysis suggested that there may be a halo effect at the level of the individual task whereby the rating of one category strongly influences an adjacent category. This effect had previously been identified in the analysis of the ratings based on the CSWE scale. There is also some evidence to suggest that there may be overlap between some of the performance criteria. If it could be established that this is the case, then some of the criteria could be collapsed or eliminated if they

were found to be redundant. However, further research into the rating process is needed to investigate raters' use of scale points and their interpretation of differing levels of performance. This would need to involve the use of introspective and retrospective methods in order to probe decision-making processes (see Smith, this volume, Chapter 5, for an example of such a study).

Implications

Developing parallel assessments

The study reported in this chapter has revealed a complex set of influences on writing task performance arising from multiple interactions between raters, tasks and rating categories, overlaid by features of the context in which assessment takes place. Given the multiple influences which affect language task performance, it would be unrealistic to expect that it would ever be possible to design tasks that were exactly equivalent. In the words of Purves (1992:110):

... there will always be task effects ... and such effects may be exacerbated by some other force — perhaps instruction, perhaps rating style or other aspects of rating difference.

Nevertheless, we can conclude on the basis of this study that some assessment tasks appear to be subject to less variability than others and are hence more likely to provide reliable information on learner outcomes. A bank of tasks therefore needs to be built up which have been carefully scrutinised, tried out with learners and discussed in detail with assessors. In this context, there may be grounds for developing a detailed set of specifications for each type of assessment task which can be used as a template by teachers who need to design their own tasks, along the lines of those suggested by Lynch and Davidson (1994). Subject to resources, new tasks could be piloted and calibrated using Rasch analysis and used to build up a library of exemplars against which new tasks could then be compared. At the same time, regular moderation sessions need to be held at which samples of student performance are carefully reviewed and reasons for rating decisions discussed at length (Claire, forthcoming). Examples of 'non-achievement' as well as successful performance need to be analysed in detail so that judges can build up a common set of interpretations and specific examples of what constitutes achievement of a given competency. This is particularly important where imprecisely worded rating criteria such as 'appropriacy' are used.

Monitoring performance criteria

Some of the performance criteria, particularly by those which use impressionistic terminology, are obviously difficult for assessors to interpret, as Smith's study (Chapter 5, this volume) also shows. However, as Davies (1992:14) comments, the problem of imprecise criteria is almost impossible to avoid unless one wants to end up with a standardised test:

The paradox is that through the attempt to refine proficiency scales by removing their defects (the imprecise and relativistic terminology-limited range, control of some structures, many error types) the precision of the descriptors tends more and more towards a list or bank of test items.

Nevertheless, those criteria on which consistent agreement cannot be reached or which appear to be redundant need to be modified or removed from the competency statements. In addition, the functioning of the performance criteria used to assess each competency needs to be subjected to ongoing research and monitoring. In this respect, analyses such as the present study can help to identify problematic criteria or inconsistent rating patterns on the basis of which such modifications can be made.

The issue of the differential difficulty of the different performance criteria also needs to be addressed. On the basis of the results of this study of these two writing competencies, it would appear that it is much easier to achieve success on some performance criteria than it is on others. This may well be true in the case of other skills. Given the large differences in the difficulty of the performance criteria, it would be useful to examine the feasibility of using differential weighting of criteria or the use of partial credit scoring.

Using multiple rating

One of the strengths of the CSWE is that the competencies (with the exception of those involving listening skills) provide a comprehensive sample of a range of language use tasks that candidates would have to engage in in everyday life. To this extent, a CSWE profile of competency achievement in writing — which might report achievement on up to four assessment tasks within the same language skill area — provides a fuller sampling than a standardised test which might contain two or three tasks at most. However, if the information on attainment of individual competencies is to be dependable, the results of this

study suggest that more than one task and one judge should be used to assess each competency. Although more resource-intensive, an approach which involves multiple tasks and raters has the potential to encourage professional exchange and discussion. At the same time, it also lends itself to profiling approaches which enable teachers to build up samples of different types of student work which reflect progress over a period of time.

Professional development

Finally, the implications of an assessment system which relies on teachers to develop and administer their own assessment tasks need to be considered. Entrusting teachers with this responsibility rests on the assumption that they have the expertise required to carry out the task effectively. However, this assumption may not be warranted. Research evidence from the general educational literature suggests that teachers receive inadequate assessment training as part of their professional preparation (Stiggins 1991, 1992; Brookhart 1994; Cizek et al 1995). In the light of this finding, it is hardly surprising that Cizek et al (1995:173) conclude that 'teachers' assessment practices do not necessarily conform to what measurement specialists would consider to be sound testing and grading practice'. There is no reason to suppose that the situation is any different as far as language teachers are concerned (Bailey and Brown 1996; Brindley 1997). In fact, the problem may be even more acute, given that language testing seems to have evolved as a parallel 'sub-profession' within language teaching and has reached such a high level of technical sophistication that, according to some commentators, it seems inaccessible (and unattractive) to many practitioners (Jones 1985; Stevenson 1985). If they are to be expected to design and conduct assessments which can provide valid and dependable information, teachers need the opportunity to develop the skills necessary to do so. While formal degree courses and professional development activities undoubtedly play an important role here, there are also other ways in which teachers can build assessment expertise — these include regular moderation sessions during which samples of learner performance are discussed (Gipps 1994; Radnor and Shaw 1995) and collaborative test development projects involving teachers and external researchers (Shohamy 1992; Brindley, 1997). However, all of these activities imply a major investment on the part of educational authorities in planning and resourcing teacher professional development.

Endnote

- 1 I would like to express my gratitude to Helen Slatyer for her assistance with data collection and collation and to Steve Ross, Brian Lynch and Tim McNamara for their advice on the various forms of statistical analysis that are reported in this chapter. I am also grateful to Donna Williams and Stephanie Claire for their organisation of the rating sessions and to Mary Kerstjens and Helen Joyce for allowing NCELTR access to facilities at NSW AMES and RMIT University. Grateful acknowledgement is also due to the 24 teachers in Sydney and Melbourne who gave up their time to participate in the project.

References

- Bachman, L F, B K Lynch and M Mason 1995a. 'Investigating variability in tasks and rater judgements in a performance test of foreign language speaking'. *Language Testing*, 12, 2: 238–57
- Bachman, L F, F Davidson and M Milanovic 1996. 'The use of test method characteristics in the content analysis and design of EFL proficiency tests'. *Language Testing*, 13, 2: 125–50
- Bailey, K M and J D Brown 1996. Language testing courses: what are they? In A Cumming, and R Berwick (eds). *Validation in Language Testing*. Clevedon, Avon: Multilingual Matters, 236–56
- Brennan, R L 1992. 'Generalizability theory'. *Educational Measurement: Issues and Practice*, 11, 4: 27–34
- Brindley, G 1997. 'Assessment and the language teacher: Trends and transitions'. *The Language Teacher*, 21, 9: 37, 39
- Brookhart, S M 1994. 'Teachers' grading: Practice and theory'. *Applied Measurement in Education*, 7, 4: 279–301
- Brown, A 1995. 'The effect of rater variables in the development of an occupation-specific language performance test'. *Language Testing*, 12, 1: 1–15
- Burrows, C 1995. 'Why assessing oral language is not a waste of time'. *Interchange*, 23: 32–4
- Christie, J and S Delaruelle 1997. *Assessment and moderation: Book 1. Task design*. Sydney: National Centre for English Language Teaching and Research, Macquarie University
- Cizek, G, S M Fitzgerald and R Rachor 1995. 'Teachers' assessment practices: Preparation, isolation and the kitchen sink'. *Educational Assessment*, 3, 2: 159–79
- Claire, S Forthcoming. Moderating CSWE assessments: Issues and practices. To appear in G Brindley (ed). *Studies in immigrant English language assessment*, vol 2. Sydney: National Centre for English Language Teaching and Research, Macquarie University
- Crick, J and R L Brennan 1984. GENOVA: *A general purpose analysis of variance system*. Version 2.2. Iowa City, IA: The American College Testing Program
- Davies, A 1992. 'Is language proficiency always achievement?' Melbourne *Papers in Language Testing*, 1, 1: 1–11
- Delaruelle, S 1997. Text type and rater decision-making in the writing module. In G Brindley, and G Wigglesworth (eds). *access: Issues in language test design and delivery*. Sydney: National Centre for English Language Teaching and Research, Macquarie University, 215–42
- Engelhard, G 1992. 'The measurement of writing ability with a many-faceted Rasch model'. *Applied Measurement in Education*, 5, 3: 171–91
- Gipps, C 1994. *Beyond testing*. London: The Falmer Press
- Haertel, E 1993. April. 'Evolving conceptions of the generalizability of performance assessment'. Paper presented at the AERA conference, Atlanta
- Hamp-Lyons, L and S P Mathias 1994. 'Examining expert judgements of task difficulty on essay tasks'. *Journal of Second Language Writing*, 3, 1: 49–68
- Hamp-Lyons, L and G Henning 1991. 'Communicative writing profiles: An investigation of the transferability of a multiple-trait scoring instrument across writing assessment contexts'. *Language Learning*, 41, 3: 337–73
- Henning, G 1987. *A guide to language testing*. Rowley, Massachusetts: Newbury House
- Henning, G 1988. 'The influence of test and sample dimensionality on latent trait person ability and item difficulty calibrations'. *Language Testing*, 5, 1: 83–99
- Jones, R 1985. Second language performance testing: An overview. In P C

- Hauptman, R LeBlanc and M B Wesche (eds). *Second language performance testing*. Ottawa: University of Ottawa Press, 15–24
- Lane, S and D Sabers 1989. 'Use of generalizability theory for estimating the dependability of a scoring system for sample essays'. *Applied Measurement in Education*, 2, 3: 195–205
- Linacre, J M and B D Wright 1993. *A user's guide to FACETS*. Chicago: Mesa Press
- Lumley, T J N, B K Lynch and T F McNamara 1994. 'A new approach to standard-setting in language assessment'. *Melbourne Papers in Language Testing*, 3, 2: 19–39
- Lynch, B K and T F McNamara 1998. 'Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants'. *Language Testing*, 15, 2: 158–80
- Lynch, B K and F Davidson 1994. 'Criterion-referenced language test development: Linking curricula, teachers and tests'. *TESOL Quarterly*, 28, 4: 727–44
- McNamara, T F 1996. *Second language performance testing: Theory and research*. London: Longman
- McNamara, T F and B K Lynch 1997. A generalizability theory study of ratings and test design in the writing and speaking modules of the access: test. In G Brindley and G Wigglesworth (eds). *access: Issues in language test design and delivery*. Sydney: National Centre for English Language Teaching and Research, Macquarie University, 197–214
- McNamara, T F and R Adams 1991. 'Exploring rater characteristics with Rasch techniques'. In selected papers of the 13th Language Testing Research Colloquium (LTRC). Princeton, NJ: Educational Testing Service
- Pollitt, A and C Hutchinson 1987. 'Calibrating graded assessments: Rasch partial credit analysis of performance in writing'. *Language Testing*, 4, 1: 72–92
- Purves, A 1992. 'Reflections on research and assessment in written composition'. *Research in the Teaching of English*, 26, 1: 108–22
- Radnor, H and K Shaw 1995. Developing a collaborative approach to modera-

- tion. In H. Torrance (ed). *Evaluating authentic assessment*. Buckingham: Open University Press, 124–44
- Shohamy, E 1992 *The power of tests: The impact of language tests on teaching and learning*. Washington, DC: National Foreign Language Center
- Stevenson, D 1985. Pop validity and performance testing. In Y P Lee, C Y Y Fok, R Lord and G Low (eds). *New directions in language testing*. Oxford: Pergamon, 111–18
- Stiggins, R 1991. 'Facing the challenge of a new era of educational assessment'. *Applied Measurement in Education*, 4, 4: 263–73
- Stiggins, R 1992. 'High quality classroom assessment: What does it really mean?' *Educational Measurement: Issues and Practice*, 11, 2: 35–9
- Vaughan, C 1991. Holistic assessment: What goes on in the raters' minds? In L Hamp-Lyons (ed). *Assessing second language writing in academic contexts*. Norwood, NJ: Ablex, 111–26
- Wigglesworth, G 1993. 'Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction'. *Language Testing*, 10, 3: 305–36
- Wright, B D and J M Linacre 1991. *A user's guide to Bigsteps Rasch-Model computer program*. Version 2.2. Chicago: Mesa Press
- Wright, B D and A Tennant 1996. 'Sample size again'. *Rasch Measurement Transactions*, 9, 4: 468

5

Rater judgments in the direct assessment of competency-based second language writing ability

David Smith

Introduction

The introduction of competency-based models of language and literacy education in Australia has, to a large degree, coincided with the return to subjective rating as the most common means of evaluating second language writing ability. However, despite the current dominance of competency-based approaches, there has been relatively little research into rater reliability in the context of competency-based assessments.

In addition, according to Hamp-Lyons (1990), much of the existing research on rater reliability and judgment in the direct assessment of writing ability has mistakenly focused on scoring procedure as the major source of diminished reliability rather than on the reader or assessor. Vaughan (1991) argues that past research has focused primarily on the *product* of raters' evaluations of writing ability and that very little is known about the decision-making *processes* and behaviours adopted by raters when making highly specialised assessment decisions. Lack of knowledge in this area not only makes it difficult to train raters; it also makes it difficult to provide direct validation of the methods and criteria used for directly assessing second language writing ability (Cumming 1990).

While some researchers have recently begun to investigate such processes within the context of holistic writing assessment, empirical accounts of the ways raters directly assess second language writing ability within competency-based models of language and literacy education are almost non-existent. It is the aim of the study presented in this chapter to contribute to this hitherto neglected field of research, with particular reference to the way in which writing assessments are conducted in the context of the Certificates in Spoken and Written English (CSWE) (NSW AMES 1998).

Reliability in competency-based assessment

The introduction of competency-based approaches to language and literacy assessment reflects a major move away from standardised language testing towards the use of performance-based assessment which directly reflects learning activities within the context in which learning takes place. Because performance assessment is criterion-referenced rather than norm-referenced, learners are not compared with each other; rather, their performance is judged according to how well it fulfils the behavioural expectations of the particular language task they are responding to. Thus, in the competency-based assessments of language performance that are used in conjunction with the CSWE, assessment decisions are made on a simple yes/no basis; that is, the assessor or rater is required to make a binary judgment on whether or not the learner demonstrates the successful attainment of a particular language task, skill or competency.

In traditional language testing, the level of agreement between raters (inter-rater reliability) is investigated by correlating the scores assigned by raters to specific learner performances on standardised tests. However, many assessment researchers have argued that traditional notions of reliability as agreement between independent judges may not be appropriate in the context of performance assessment or competency-based assessment since these types of assessment are not designed to emphasise difference among learners. Nor is there an underlying scale of performance based on 'true scores' (Gipps 1994). As a consequence, some researchers maintain that reliability needs to be reconceptualised where performance assessments are concerned. Gipps (1994), for example, proposes that in cases where traditional reliability measures are not appropriate, we should discard the term *reliability* altogether and instead replace it with *comparability* which is based on consistency. According to Gipps (1994:171):

Consistency leading to comparability is achieved by assessment tasks being presented in the same way to all learners being assessed; assessment criteria being interpreted in the same way by all teachers; and learner performance being evaluated according to the same rubric and standards by all markers.

Such a paradigm shift has significant methodological implications for it strongly suggests that in order to achieve consistency of assessment within competency-based models of language and literacy education, we need to

know more about the raters themselves, the nature of reading for the purpose of assessment and the process of reading according to specific assessment guidelines. As Huot (1990:258) observes:

It seems that if assessment literature is to progress, more inquiries are needed about how raters arrive at judgments about writing quality and what part rating procedures have in this process.

Research on rater judgments

One approach to the investigation of the processes through which raters arrive at their assessment decisions is through the use of raters' think-aloud protocols or verbal reports. In this model of data collection, raters verbalise their thoughts while reading and assessing written texts and their verbalisations are recorded, transcribed and coded according to either an a priori scheme or a coding scheme developed through a preliminary analysis of the data (Weigle 1994). Some reservations have been expressed in the literature, however, about whether think-aloud verbal reports can accurately reflect or capture the cognitive processes that they are designed to uncover. For instance, Nisbett and Wilson (1977) have argued that a verbal report is not a complete record of a person's thought processes and question whether this method can provide introspective access to higher cognitive processes. They make the point that the absence of any particular phenomenon in a protocol may not be evidence of its absence in actuality, since it is not possible to report on all of one's thoughts at any given moment. On the other hand, Ericsson and Simon's (1980) thorough review of the literature on tracing cognitive processes concluded that think-aloud protocols, when elicited with care and interpreted with full understanding of the circumstances under which they were obtained, are a valuable and thoroughly reliable source of information about cognitive processes.

A number of studies have used think-aloud protocols to investigate rater decision-making behaviour in the holistic assessment of second language writing ability. Cumming (1990), in one of the earliest studies of rater behaviour in the context of second language writing assessment, used think-aloud protocols in order to describe the decision-making behaviours used by experienced and inexperienced raters when assessing second language writing ability holistically. Cumming analysed the think-aloud protocols of seven expert and six novice raters rating 12 ESL essays selected from a pool of examination papers administered as a placement test for ESL classes at a Canadian university. The essays

were selected to represent two levels of ESL proficiency (intermediate and advanced); two levels of writing expertise in students' mother tongue (professional and average); and different language and cultural backgrounds of students. The essays were evaluated holistically on language use, content and organisation. Through an impressionistic descriptive analysis of the raters' verbal reports, Cumming identified 28 decision-making behaviours used by raters to interpret and evaluate the student essays. Cumming classified these behaviours into two types of strategies: 'interpretive strategies' used to read the texts, and 'judgment strategies' used to evaluate the qualities of the texts. The importance of Cumming's study lies in the fact that it is among the first of its type to describe, albeit in a preliminary way, the decision-making behaviours which raters perform mentally while directly assessing second language writing ability. It is also important because it lays the foundation for a model of the thinking processes integral to the skill of writing assessment.

Vaughan (1991) also used think-aloud or verbal protocols to study rater behaviour in second language writing assessment. She analysed the think-aloud protocols of nine experienced judges rating six ESL compositions on a six-point holistic scale. From her analysis, Vaughan identified four approaches to the holistic assessment of writing ability: 'the first-impression-dominates approach'; 'the single-focus approach'; 'the two-category strategy'; and 'the grammar-oriented' rater. Vaughan found that raters focused on a number of key linguistic features or elements when assessing second language writing ability: *content, organisation and grammar*, followed by *handwriting and punctuation*. This conclusion would appear to support the findings of research on first language writing assessment which also concluded that raters are primarily concerned with content and organisation.

On the basis of these findings, Vaughan argued that despite their similar training, different raters focus on different essay elements and have distinctly individual approaches to reading essays. Vaughan suggests that while raters can agree on many essays based on the guidelines for holistic assessment, they may fall back on their own rating style for essays which do not clearly fit the descriptors of the holistic scale. The results of Vaughan's study imply that raters are not adhering to a single, internalised method for judging second language writing ability. In addition, it appears that despite the use of assessment guidelines, raters tend to rely on their own rating strategies when assessing writing samples which are deemed to be 'borderline' cases.

A more recent study of the decision-making behaviours of composition markers in the context of second language writing assessment is provided by Milanovic, Saville and Shuhong (1996). Like Cumming and Vaughan, Milanovic et al also used think-aloud protocols in order to investigate raters' decision-making behaviour when rating ESL compositions using holistic techniques. They analysed the think-aloud protocols, retrospective written reports and group interviews of 16 experienced judges rating 40 ESL essays on a five-point holistic rating scale. Of these essays 20 were selected from an intermediate level ESL examination and 20 were selected from an advanced level examination. An impressionistic analysis of the data revealed that raters employed four identifiable rating strategies: 'the principled-two scan/read'; 'the pragmatic-two scan/read'; 'the read-through'; and 'the provisional-mark'. In addition, Milanovic et al identified 11 linguistic features that raters claimed to focus on when rating: *length, legibility, grammar, structure, communicative effectiveness, tone, vocabulary, spelling, content, task realisation and punctuation*. Although the researchers were unable to obtain any exact data regarding the effect of these linguistic elements on the final scores, they were able to provide some insights into the possible weight attributed to each element while rating. In terms of grammatical features, for instance, while raters often spoke about mistakes 'impeding meaning', it seemed that the majority of raters sought to balance grammatical accuracy with communicative competence, giving the latter a slight priority. The most subjective compositional element identified by Milanovic et al was *content*. In relation to this element, Milanovic et al noted that raters' personal responses to the content of a script most often influenced a students' final score. This finding supports previous research which has also found that raters were most influenced by the content of a written script when rating holistically (Freedman 1979).

Rater consistency in competency-based assessment

The research literature discussed thus far has been based on the use of holistic scoring procedures in the direct assessment of second language writing ability. However, competency-based models of language and literacy education, like that represented by the CSWE, do not employ such methods of assessment. Only recently have researchers begun to investigate the reliability of competency-based second language assessment procedures.

Unpublished research conducted by Jones (1993) into rater consistency in the assessment of competencies at Certificate II level of the CSWE indicated high

levels of overall rater agreement for most competencies. While this finding is not particularly informative in itself, Jones' study does, however, provide some useful data on the interpretation of performance criteria. Jones found, for instance, that some of the relativistic terminology used in the performance criteria statements (eg 'mostly appropriate'; 'mostly correct spelling') was difficult for raters to interpret.

Burrows (1994) conducted a pilot study to examine the reliability of the assessment of two written competencies at Certificate II level: Competency 10 (*Can write short reports relevant to further education*) and Competency 11 (*Can write short essays relevant to further education*). Data was collected from seven teachers from six different Adult Migrant Education Services (AMES) regions and a total of 77 texts were examined. Using a coding system based on a systemic functional view of grammar, Burrows statistically analysed two features of the written texts: correct and incorrect uses of English at the discourse level and incorrect uses of English at the lexico-grammatical level. On the basis of her analysis, Burrows found that the teachers rating the written texts were responding to the same features within the texts to make their assessment decisions. In other words, raters were referring to the prescribed performance criteria statements when making their assessments. The results also indicated that reliable assessment occurred, particularly at the lexico-grammatical level. On the basis of these findings, Burrows concluded, at least for the two written competencies examined, that there was a high degree of rater reliability.

Brindley (forthcoming) investigated the ratings awarded by a group of 12 trained CSWE assessors to 72 written texts which had been submitted by AMES teaching centres as 'benchmark' examples of the competency in question. He found that the assessors did not consistently agree that the benchmark texts were in fact examples of minimal competency achievement. The assessors also showed a good deal of variability both in their interpretation and application of individual performance criteria.

Quantitative analysis of ratings revealed that the performance criteria were of a different order of difficulty. Brindley concluded that careful attention needs to be paid to the way in which the tasks are specified in order to ensure that the relevant features of performance are elicited. He also suggested that some performance criteria may be superfluous or inappropriate to the task at hand and may need to be revised.

Aims and objectives

The brief overview of research reported above suggests that, despite recent developments in the field of holistic assessment, there remains a distinct lack of qualitative process data about readers as they 'make decisions about written products and shape their judgments to the parameters of the assessment instrument that they are applying' (Hamp-Lyons 1990). Nowhere is this lack of qualitative data more obvious than in relation to competency-based second language writing assessment.

As a contribution to the line of research undertaken by Cumming, Vaughan and Milanovic et al, this study therefore sets out to investigate the decision-making strategies and behaviours adopted by raters when directly assessing second language writing ability within the context of a competency-based curriculum and assessment framework.

Specifically, the study aims to:

- Examine the degree to which the criteria used to assess written competency within Certificate II of the CSWE can ensure acceptable levels of rater consistency.
- Describe and investigate the decision-making strategies and behaviours adopted by raters when assessing competency-based second language writing ability.

The following research questions will be addressed:

- Do the criteria which have been designed to directly assess second language writing ability within CSWE II ensure acceptable levels of rater consistency?
- How do raters interpret and apply assessment criteria when assessing written texts?
- What reading strategies and behaviours do raters exhibit when assessing written texts?
- What features of written texts do raters focus on when making assessment judgments?

Methodology

Setting and subjects

Silverlake English Language Centre provided the setting for the study.¹ This centre is located in a large culturally diverse suburb in metropolitan Melbourne. It is funded by the Federal Government through the Adult Migrant English Program (AMEP) to provide specialist language and literacy tuition for on-arrival/settlement students and is nationally accredited to deliver the CSWE.

The six informants/raters selected for the study all taught at the Silverlake English Language Centre and were employed by Adult Multicultural Education Services (AMES) Victoria. Informants were all chosen on the basis of their experience in assessing second language writing ability within the CSWE and on their interest in second language assessment issues. All informants had been trained in the principles of competency-based second language assessment and all had participated in regular CSWE assessment moderation sessions.

Data collection

Each rater was first asked to assess the same three written texts according to the prescribed CSWE performance criteria for that competency — in this case, Competency 14 (*Can write a short recount*). (See Figure 4). The three texts were then photocopied and given to each rater in exactly the same order and at exactly the same time as it was felt that sequencing the texts differently may introduce inconsistency and affect rating behaviour. In addition, the texts were not altered in any way. Following the think-aloud procedures described by Ericsson and Simon (1980), raters were then asked to give a verbal report on their assessments for each written text. It was made explicit to raters that the report should consist of think-aloud, stream-of-consciousness disclosure of thought processes while assessing and should be unedited and unanalysed. Raters were instructed not to switch the tape recorder off until they had completed their assessments. All raters recorded their assessments individually in the interview room of the Silverlake English Language Centre during the normal course of the working day.

Analysis

The recorded think-aloud verbal protocols were transcribed in full and then reviewed impressionistically to develop a coding scheme. The transcripts were analysed in a number of stages. First, the transcriptions of each individual

Elements	Performance Criteria	Range Statements	Evidence Guide
Discourse Structure	<ul style="list-style-type: none"> i. can use appropriate staging ii. can use appropriate conjunctive links iii. can use simple reference appropriately 	<ul style="list-style-type: none"> • uses appropriate temporal staging • uses some conjunctive links eg 'first', 'then', 'and', 'but' • uses simple reference appropriately eg pronouns and articles 	<ul style="list-style-type: none"> • familiar and relevant topic • approximately 100 words in length • recourse to dictionary • may include a few grammatical and spelling errors but errors do not interfere with meaning
Grammar and Vocabulary	<ul style="list-style-type: none"> iv. can use vocabulary appropriate to the topic v. can use past tenses and other past markers vi. can construct 2 clause sentences 	<ul style="list-style-type: none"> • uses vocabulary appropriate to the topic • uses past tenses and other past markers • constructs some sentences containing 2 clauses 	<ul style="list-style-type: none"> • Learners write about a past event eg excursion, workplace visit, holiday, picnic, experience story.

Figure 4: CSWE Writing Competency 14

Graphology

Specific performance criteria related to graphology have not been included. However it is assumed that:

- there may be some inaccuracies in letter formation, spelling of multi-syllabic words, layout and punctuation
 - teaching programs will pay attention to graphological features and to self-monitoring and self-correction strategies
- In CSWE II the punctuation focus will be on capital letters, fullstops, question marks, commas and inverted commas.

Source: NSW AMES 1995

raters' think-aloud verbal reports for each of the three texts assessed were analysed in order to determine the degree of consistency with which raters classified the text and the performance criteria. For both the purposes of this study and for the purposes of CSWE recording and reporting requirements, 'A' or Achieved represented a pass and 'P' or Partially Achieved represented a fail.

Second, the transcriptions of the think-aloud verbal reports were read through, categorised and analysed a number of times to identify the ways in which individual raters interpreted and applied the performance criteria statements. 'Interpretation' was defined as the way in which key terms in the performance criteria statements were defined and operationalised by raters. 'Application' was defined as the way in which raters' interpretations of the performance criteria statements were applied to their assessment of the texts. Once these individual interpretations and applications had been identified, they were then compared with each other in order to investigate whether or not raters were interpreting and applying the performance criteria in similar ways.

Next, the individual think-aloud verbal reports were analysed holistically in order to identify the reading strategies or styles used by raters when making their assessments. In order to identify raters' reading strategies, each rater's think-aloud verbal reports were analysed for comments and patterns of response that suggested that they were reading in different ways.

Finally, the transcripts were analysed in order to identify the textual features that raters commented on when making their assessment judgments. A 'comment' was defined as a meaningful unit in the form of a phrase, a word, a clause or a whole sentence. These units were identified and frequency counts made. Following this, the comments were grouped into nine general categories based on both the categories identified in previous research and those that emerged from the data analysis.

Results

The presentation of the results will be structured around four key questions:

- 1 Do the criteria which have been designed to directly assess second language writing ability within the Certificates in Spoken and Written English II result in consistency of assessment?
- 2 How do raters interpret and apply the assessment criteria when assessing written texts?

- 3 What reading strategies do raters exhibit when assessing written texts?
- 4 What textual features do raters focus on when making their assessment judgments?

Rater consistency

Table 29 shows that the overall level of rater consistency was relatively high; that is, the six raters generally were in agreement as to whether the texts achieved the competency or not.

Table 29: Rater consistency in text classification

	Maria	Bill	Natasya	Dean	Andrew	Anne
Text 1						
Criterion i	P	A	A	A	P	A
Criterion ii	A	P	P	A	P	A
Criterion iii	A	A	A	A	A	A
Criterion iv	A	A	A	A	A	A
Criterion v	A	A	A	A	A	A
Criterion vi	P	A	P	P	P	A
Assessment	P	A	P	P	P	A
Text 2						
Criterion i	A	A	A	A	A	A
Criterion ii	A	A	A	P	A	A
Criterion iii	A	A	A	A	A	A
Criterion iv	A	A	A	A	A	A
Criterion v	A	A	A	A	A	A
Criterion vi	A	A	A	A	A	A
Assessment	A	A	A	P	A	A
Text 3						
Criterion i	A	A	A	A	A	A
Criterion ii	A	A	A	A	A	A
Criterion iii	A	A	A	A	A	A
Criterion iv	A	A	A	A	A	A
Criterion v	A	A	A	A	A	A
Criterion vi	P	A	A	A	A	A
Assessment	P	A	A	A	A	A

A = achieved/ P = partially achieved

As Table 29 shows, Text 1 was judged to have partially achieved the competency by four of the six raters while Text 2 and Text 3 were both judged to have achieved the competency by five of the six raters.

In terms of individual patterns of response to the assessment of the three texts, Table 1 indicates that Bill and Anne judged all three texts to have achieved the competency, Natasya and Andrew judged two of the texts to have achieved the competency and Maria and Dean judged one of the texts to have achieved the competency. There was less agreement among raters, however, at the level of individual performance criteria, particularly in relation to the assessment of Text 1, in terms of both the total number of performance criteria achieved and the type of performance criteria achieved.

Of the three texts assessed, Text 1 was the most problematic in terms of rater consistency. Not only was there a lack of agreement among raters as to whether or not Text 1 achieved the competency overall, there was also lack of agreement at the level of individual performance criteria. The six raters were in substantial disagreement in relation to the number and type of performance criteria that the text was able to successfully demonstrate. In other words, there was no agreement among raters as to why the text failed to achieve the competency.

Raters' reading strategies

Analysis of the data revealed that three identifiable reading strategies or styles were used in the assessment process. These will be referred to as the 'read-through-once-then-scan' approach, the 'performance criteria-focused' approach, and 'the first-impression-dominates' approach. Each approach is discussed below.

'Read-through-once-then-scan' approach

The first reading strategy identified, the 'read-through-once-then-scan' approach, was adopted by two raters, Maria and Dean. Raters adopting this strategy read through the text once. This initial reading of the text was conducted without a break in the reading flow, without comment and without judgment. Raters then scanned the text for evidence of the successful demonstration of each of the performance criteria statements. The 'read-through-once-then-scan' approach can be illustrated by the following comments.

Maria: *I'm going to read Text 1 ... The student uses past tense correctly.*

We have [scanning the text] went, arrived, took, finished, came, cooked, played yeah that's correct.

Dean: *Um okay, reading through it first ... Uses some conjunctive links. Um just glancing over it, yeah there's and, my wife bought a jacket gave for me and I bought a jumper gave for my wife.*

'Performance criteria-focused' approach

The second reading strategy identified, the 'performance criteria-focused' approach, was adopted by Natasya and Andrew. Raters adopting this strategy did not read through the text in its entirety, rather they scanned the text for evidence of the successful demonstration of the performance criteria. Once sufficient evidence or examples were located, the reading process was terminated and scanning for evidence of the next performance criteria statement began. The 'performance criteria-focused' approach can be seen in the following rater comments:

Natasya: *Text 2. Can write a short recount. Write about your weekend. Criterion i: Can use appropriate staging. The writer organises all ideas in three paragraphs. The first sentence is the beginning stage. Last weekend I went shopping.*

Andrew: *Text 3. Um, uses appropriate temporal staging. Last weekend I had a weekend and very happy. Ah, Last weekend I had a weekend. I don't know about that. Okay.*

'First-impression-dominates' approach

The third reading strategy identified, the 'first-impression-dominates' approach was adopted by Bill and Anne. Raters adopting this strategy read through the text once, commented on whether or not the text achieved the competency and then scanned the text for evidence of the successful demonstration of each of the performance criteria statements in order to vindicate their initial judgment. The 'first-impression-dominates' approach is illustrated by the following comments.

Bill: *Just looking through this just to make sure that it's all correct and I can follow the gist because ... it's important that I'm able to actually follow it and get the general gist. So while I've done that I think at this stage my gut feeling is that it um has passed.*

Anne: Before I do it I just want to read them to see what the sort of levels are like, and in relation to her first impression of Text 1, That's not bad actually, that's okay.

There were, however, variations between the two raters using this strategy. Bill, for instance read through the text without a break in the reading process, whereas Anne interrupted the reading process to comment on or correct errors.

Analysis of the data on rater reading strategies suggested that a relationship appeared to exist between the type of reading strategy adopted by raters and the number of texts judged to have achieved the competency. The results of this analysis are presented in Table 30.

Table 30: Relationship between reading strategy and number of texts passed

Rater	Reading strategy	Number of texts passed
Maria	Read through/scan	1 (Text 2)
Bill	First impression	3 (Text 1/2/3)
Natasya	Criteria focused	2 (Text 2/3)
Dean	Read through/scan	1 (Text 3)
Andrew	Criteria focused	2 (Text 2/3)
Anne	First impression	3 (Text 1/2/3)

As Table 30 indicates, raters adopting the 'first-impression-dominates' approach to reading passed the greatest number of texts, whereas raters using the 'read-through-once-then-scan' approach passed the least number of texts. In terms of the relationship between degree of rater consistency and the type of reading strategy employed, Table 30 indicates that, with the exception of raters using the 'read-through-once-then-scan' approach, raters adopting the same reading strategy were in full agreement with each other in terms of the number and type of texts they passed or failed.

Textual features identified by raters

Even though raters made their assessment judgments on the basis of the extent to which the texts met the performance criteria, they nonetheless commented on a range of textual features extraneous to the performance criteria. As Table 31 shows, grammar, organisation and coherence were the textual features commented on by the greatest number of raters.

Table 31: Textual features commented on by raters extraneous to performance criteria

Feature	Number of raters who commented
Grammar	5 (Maria, Bill, Dean, Andrew, Anne)
Organisation	4 (Bill, Natasya, Dean, Anne)
Coherence	3 (Bill, Andrew, Anne)
Sentence structure	2 (Maria, Natasya)
Punctuation/capitalisation	2 (Maria, Andrew)
Spelling	2 (Maria, Anne)
Handwriting	2 (Bill, Anne)
Length of text	2 (Bill, Anne)
Lexical choice	2 (Bill, Natasya)

While extraneous to the performance criteria statements, five of the nine features commented on, *grammar, punctuation and capitalisation, spelling, handwriting and length of text*, are explicitly stated within the competency's range statements. Combined with the performance criteria, elements and evidence guide, the range statements complete the competency description. The four textual features commented on by raters which are not included in the competency description are those related to *organisation, coherence, sentence structure and lexical choice*. In other words, just under half of the textual features commented on by raters were absent from the assessment rubric.

Table 32 suggests that there may be a relationship between the type of reading strategies adopted by raters and the number and type of extraneous textual features commented on.

Table 32: Relationship between extraneous textual features commented on and rater reading strategy

Rater	No. of features commented on	Feature most often commented on	Reading strategy
Maria	4	Punctuation	Read through/scan
Bill	6	Coherence	First impression
Natasya	3	Sentence structure	Criteria focused
Dean	2	Grammar	Read through/scan
Andrew	3	Grammar	Criteria focused
Anne	6	Coherence	First impression

Bill and Anne, the two raters adopting the 'first-impression-dominates' approach to reading, commented on more extraneous textual features than other raters. On the other hand, Natasya and Andrew, the two 'performance criteria-focused' readers commented on the least number of textual features. In terms of the type of extraneous textual features commented on, analysis of the data suggests that with the exception of the two 'first-impression-dominates' readers, raters tended to display individual patterns of response.

Discussion

Rater consistency

The overall relatively high level of rater consistency for two of the three texts compares favourably with the levels of rater agreement that have been reported in studies using holistic methods of assessing writing ability. This would appear to suggest that competency-based performance assessments of second language writing ability are able to achieve comparable and, in some instances, higher levels of rater consistency than holistic scoring procedures. Such comparisons, however, are misleading for two reasons. First, raters assessing writing ability within a competency-based assessment framework are only required to make a binary yes/no judgment on whether or not a text achieves a competency, whereas holistic raters must assign scores. As a consequence, there is a greater margin for rater disagreement in holistic scoring based on the range of scores that can be assigned to a text. Second, overall levels of agreement may obscure differences at the level of individual performance criteria, particularly on those texts which are judged to have not achieved the competency. Caution must therefore be exercised when making comparisons between these two methods of assessment.

The results do, however, compare favourably with past research on inter-rater reliability in the assessment of competencies within the CSWE. Burrows (1994) and Jones (1993) also reported high levels of overall rater reliability, particularly in the assessment of written competencies, at both Certificate II and Certificate III. However, according to Brindley (1994:52), since agreement is only deemed to occur when all of the performance criteria have been met, overall levels of agreement give only a partial picture. Of perhaps greater interest is the degree of rater agreement at the level of individual performance criteria.

The finding that there was a high degree of overall rater consistency in terms of whether the texts achieved the competency or not obscures the fact that there

was considerably less agreement, particularly in relation to Text 1, at the level of individual performance criteria. For example, while four of the six raters judged Text 1 to have failed to achieve the competency there was no agreement among raters in terms of the number and type of performance criteria the text failed to demonstrate.

At the level of individual performance criteria, the most problematic criteria at the discourse structure level were Performance Criterion i (*Can use appropriate temporal staging*) and Performance Criterion ii (*Can use some conjunctive links*) and, at the lexico-grammatical level, Performance Criterion vi (*Can construct some sentences containing two clauses*).

The finding that there was generally less agreement among raters at the discourse structure level than at the lexico-grammatical level, particularly in relation to Text 1, is consistent with the results reported by Burrows (1994). Analysis of raters' think-aloud verbal reports suggests, however, that these differences may be due not so much to the nature of the productive written language skills being assessed, but to the way in which raters interpret and apply the individual performance criteria statements. These differences are discussed further below.

Interpretation and application of Performance Criteria

Analysis of raters' think-aloud verbal reports suggests that individual raters interpret and apply performance criteria in a number of different ways. In situations where the text being assessed does not unambiguously demonstrate the achievement of a particular performance criterion, these differing interpretations may lead to low levels of rater consistency.

A number of performance criteria appeared to be particularly subject to varying interpretations. These were:

- Performance Criterion ii: *Can use some conjunctive links*, eg **first, then, and, but**
- Performance Criterion iii: *Can use simple reference appropriately*, eg **pronouns** and **articles**
- Performance Criterion iv: *Can use vocabulary appropriate to the topic*
- Performance Criterion v: *Can use past tense and other past markers*.

Each of these criteria seemed to evoke different meanings for raters. Some raters strictly applied the specific examples of conjunctions provided in

Performance Criterion ii to the texts, whereas others did not. In their application of Performance Criterion iii, some raters chose to assess both the use of pronouns and articles while other raters assessed only one of these forms of reference. For some raters the use of key words — usually nouns or noun phrases — was evidence of achievement of Performance Criterion iv, while for others it was taken for granted. Finally, in their application of Performance Criterion v, some raters assessed the use of both past tense and past markers, whereas other raters focused their assessment on only one of these, usually the use of past tense. Despite these differences of interpretation, however, there was full agreement among raters that Performance Criteria iii, iv and v had been fulfilled, as noted above.

In relation to Performance Criterion ii (*Can use some conjunctive links, eg first, then, and, but*), there was some concern expressed as to how strictly the examples given were to be applied to the text being assessed. For instance, Bill commented:

Not a lot of first, then and buts, only one or two throughout and mostly in the first paragraph. So I don't know. Do you have to have first, then and but?

Those raters who strictly applied the examples to the text, Bill and Natasya, failed those texts which were unable to demonstrate their use but who were able to use other conjunctive devices. On the other hand, the examples provided for Performance Criterion iii (*Can use simple reference appropriately, eg pronouns and articles*) appeared to impact positively on the degree of rater consistency. Some raters required texts to demonstrate the use of both pronouns and articles, whereas other raters were satisfied with the use of pronouns alone. As a consequence, there was complete agreement amongst all raters that the three texts assessed satisfied this performance criterion.

The three performance criteria statements which were most problematic in terms of rater consistency all contained terms that were interpreted differently by raters. These were:

- Performance Criterion i: *Can use appropriate temporal staging*
- Performance Criterion ii: *Can use some conjunctive links, eg first, then, and, but*
- Performance Criterion vi: *Can construct some sentences containing two clauses.*

In relation to Performance Criterion i, raters appeared to interpret the term 'temporal' in least three different ways. Bill stated that:

I don't really like this because I'm not really sure what this is um temporal, I'm not sure.

Natasya and Dean interpreted 'temporal staging' as meaning that the text must display a beginning, a middle and an end. On the other hand, Maria and Andrew interpreted it broadly as the logical ordering of events, while Bill and Anne interpreted it in terms of correctly sequenced prepositional time phrases. Despite this lack of consistency, however, both Text 2 and Text 3 were deemed to satisfy the requirements as far as temporal staging was concerned.

It is perhaps no coincidence that two of the three most problematic performance criteria, Performance Criteria ii and vi, contained the determiner 'some'. As Anne commented in her assessment of Text 2:

Whatever some is I don't know. How many do you want?

Raters clearly had different ideas concerning the number of obligatory features of writing which had to be demonstrated. For instance, in relation to Text 1, in their assessments of Performance Criterion ii (*Can use some conjunctive links*), Natasya noted that 'only once did he use a conjunctive link in a sentence', whereas Dean observed that 'this person has three so I'll say that's achieved'.

The finding that raters interpreted the term 'some' differently is consistent with studies by Alderson (1991) and Jones (1993), both of whom found that such relativistic terminology was difficult for raters to interpret.

It is interesting to note the effect that the inclusion of 'some' had on the prescriptive nature of the performance criteria statements and, in turn, the effect that this had on rater consistency. It would appear, paradoxically, that the more prescriptive performance criteria statements were, the less agreement there was among raters. This finding is consistent with claims by Brindley (1998) who argues that even in cases where the assessment criteria are clearly specified, rater consistency remains a problem. On the other hand, Gipps (1994) suggests that the greater the amount of structure in performance assessment marking schemes, the more likely it is that raters will agree on the result. However, the possibility exists, as the results of this study suggest, that agreement among raters can conceal fundamental differences.

Despite the fact that raters interpreted and applied the performance criteria

statements in different ways, there was no evidence to suggest that raters were operating from their own personalised constructs and applying their own criteria in spite of, or in addition to, those they had been given. When making their assessment decisions, raters did not appear to refer to criteria which were not contained in the competency description. Nor did they appear to focus more heavily on the assessment of some criteria at the expense of others. This finding contradicts the claims of Brindley (1991) who argues that there is a tendency for raters to operate with their own criteria, irrespective of the criteria provided.

As the analysis suggests, raters adhered strictly to the performance criteria statements provided in making their assessments. Again, this trend is not supported in the literature. Caulley et al (1988), for instance, found that in making their assessments raters rarely referred to the assessment criteria at all. This contrasts sharply with the results of this study which found that raters constantly referred to the performance criteria statements when making their assessment judgments. In addition, analysis of raters' think-aloud verbal reports clearly demonstrated that when making their assessment judgments, raters justified their decisions with specific reference to concrete textual features in language consistent with the application of the specified assessment criteria.

It would appear, therefore, that the inconsistent ratings are attributable to raters' differing understanding of imprecisely defined terms such as 'temporal staging' and 'some conjunctive links', leading to a lack of agreement as to whether these specific performance criteria had been met. If this is so, then higher levels of rater consistency might be achievable if 1) they possessed a shared understanding of a small number of problematic terms which appear in a few performance criteria or 2) terms such as 'some' were simply removed from the assessment rubric altogether.

Raters' reading strategies

While raters assessed the texts according to the prescribed performance criteria statements, analysis of raters' think-aloud verbal reports revealed that raters adopted individual styles to the process of reading for the purposes of assessment. Three distinctive reading styles were identified: the 'read-through-once-then-scan' approach; the 'performance criteria-focused' approach; and the 'first-impression-dominates' approach. With the exception of the 'performance criteria-focused' approach, these reading strategies have previously been identi-

fied in the literature on holistic writing assessment. The 'performance criteria-focused' approach appears to be a previously unidentified reading strategy.

The finding that raters adopt individual approaches to the process of reading for assessment has considerable empirical support in the literature. Vaughan (1991) and Milanovic et al (1996) in their studies of rater reading strategies/styles in the assessment of second language writing ability using holistic techniques also found that raters exhibited characteristic strategies or styles of reading that stemmed from individual approaches to the process of assessment. Given the fundamental differences between holistic scoring and competency-based assessment, it was not expected that any commonality of reading style would exist between raters using these distinct forms of writing assessment. However, the reading strategy identified by Vaughan (1991) and Milanovic et al (1996) as the 'first-impression-dominates' approach or the 'read-through-once-then-scan' approach was exhibited by two of the raters in this study, Bill and Anne. While this approach to reading was characterised by a single reading of the text in order to gain an overall impression of its quality, the raters who adopted this approach still adhered to the performance criteria statements in order to arrive at their assessment decision. However, the performance criteria appeared to have less of a controlling influence on the reading behaviours of these raters than on raters who adopted other reading strategies.

The two raters who adopted the 'first-impression-dominates' approach, Bill and Anne, while not rating holistically, did, however, take a significantly more global view not only of the assessment process itself, but also of the possible consequences of their assessment decisions than did raters who adopted different reading styles. No notable differences were identified between raters who adopted the 'read-through-once-then-scan' approach and raters who used the 'performance criteria-focused' approach to reading. In fact, raters adopting these two reading strategies were exclusively focused on the performance criteria and their assessment decisions appeared to have been made on the basis of the criteria alone. Differences did exist, however, between those raters adopting the 'first-impression-dominates' approach to reading and those raters using other strategies. Analysis of the data revealed that raters adopting the 'first-impression-dominates' approach differed from the other raters in five significant areas: the number of texts passed, the application of the performance criteria statements, the number and type of textual features commented on, the comparison of texts, and the consequences/implications of assessment decisions.

In terms of the number of texts judged to have achieved the competency, the results of this study suggest that a relationship existed between reading strategy and the number of texts passed. The two raters who judged all three texts to have achieved the competency both adopted the 'first-impression-dominates' approach, whereas the two raters who adopted the 'performance criteria-focused' approach both passed two texts and the two raters who adopted the 'read-through-once-then-scan' approach both passed one text. This finding has significant implications for the training of raters because, although extremely tentative, it implies that some reading strategies or styles may result in higher levels of rater consistency in the direct assessment of writing ability than others. It also implies, perhaps more significantly, that even though some reading strategies may result in greater levels of agreement between raters, such approaches to reading may not be consistent with the underlying principles of competency-based assessment.

In terms of the interpretation and application of the performance criteria statements, the type of reading strategy adopted by raters did not appear to affect raters' interpretations; however, patterns of rater response indicated that in relation to Performance Criterion v (*Can use past tense and other past markers*), a clear relationship existed between the reading strategy adopted by raters and the application of this performance criteria statement. Raters adopting the 'first-impression-dominates' approach, Bill and Anne, consistently ignored the use of past markers in their assessment of this performance criteria across the three texts. On the other hand, the 'read-through-once-then-scan' approach raters, Maria and Dean, consistently assessed both the use of past tense and the use of past markers. This suggests that the 'first-impression-dominates' approach raters do not apply some of the performance criteria statements as strictly as raters using the 'read-through-once-then-scan' approach. To a certain extent, it might be expected that the 'first-impression-dominates' approach raters would adhere less strictly to the application of the performance criteria, compared to other raters, given the nature of this reading strategy. It might also be expected that raters adopting the 'performance criteria-focused' approach to reading would apply the performance criteria statements more strictly than raters adopting other reading strategies. However, analysis of the data revealed that this was not the case. In fact, there was very little agreement between the two raters using the 'performance criteria-focused' approach in terms of their application of the performance criteria. Again, we may speculate that, with the exception of the 'first-impression-dominates' raters, irrespective of the reading

strategy adopted, raters displayed highly individualistic approaches to the interpretation and application of the performance criteria statements.

Textual features identified by raters

In relation to the number and types of comments relating to textual features made by raters adopting different reading strategies, analysis of the data clearly revealed that raters using the 'first-impression-dominates' approach commented on a greater number of textual features than did other raters. Of the nine textual features commented on by raters, the two raters using the 'first-impression-dominates' approach, Bill and Anne, both commented on six features. On the other hand, raters using the 'performance criteria-focused' approach, Natasya and Andrew, both commented on only three textual features. In addition, analysis of the data on the categories of textual features raters commented on indicates that raters using the 'first-impression-dominates' approach made more comments about textual coherence than any other category, whereas raters adopting other reading strategies generally confined their comments to lexico-grammatical features. This suggests that raters using the 'first-impression-dominates' approach to reading are possibly influenced by a greater number of textual features extraneous to the performance criteria statements, particularly at the discourse structure level, than are raters adopting different approaches. However, it is important to note that raters' comments do not constitute unequivocal evidence of influence and it was not possible to obtain any insights regarding the effect these extraneous textual features had on assessment decisions. Suffice to say that reference to these features appeared to be interpretive or observational rather than judgmental.

Despite the fact that assessment within the CSWE is criterion-referenced, that is, texts are assessed against specific criteria and not against each other, both Bill and Anne compared the standard and quality of the texts that they assessed. Bill, for instance, in his assessment of Text 2 commented that: 'I think it's better than the first one.' Similarly, in her assessment of Text 3, Anne commented that: 'The other two seem to flow better.' This apparent tendency for raters adopting the 'first-impression-dominates' approach to compare texts may, however, be a function of the strategy itself rather than of any individual idiosyncrasy.

Finally, both Bill and Anne commented on the consequences of their assessment decisions in terms of promoting learners to the next CSWE certificate level

based on their assessment of a single text. Bill, for instance, in his assessment of Text 1 commented:

Do you pass a student and put him into Certificate III or do you keep them back because not every aspect is met? I think you have to take a global view.

On the other hand, Anne made the following observation:

I would call Text 1 an A- which means yes he has met all the things but I don't think it's good enough for Certificate II.

For Bill the concern was that while a text may not have achieved all of the performance criteria, it was unfair not to promote a learner to the next certificate level when, based on a 'first-impression-dominates' approach to reading, the text had satisfied the perceived standard required for learners at this certificate level. On the other hand, for Anne the concern was that while a text may have achieved all of the performance criteria, it was still not of the standard required for promotion to a higher certificate level. None of the other raters made any comments about the possible consequences or implications for learners of their assessment decisions.

This analysis suggests that raters adopting the 'first-impression-dominates' approach to reading bring to the assessment process an internalised and personalised view of what constitutes an acceptable quality or standard of writing that may be impervious to control, despite the existence of very specific performance criteria. This view, related to the issue of rater expectation, is consistent with the position taken by Huot (1990) who has argued that writing assessment procedures create a tension between the control necessary to achieve rater consistency and the natural variability present in the fluent reading of raters who possess a range of expectations as readers. The fact that the performance criteria appeared to have a diminished controlling effect on those raters adopting the 'first-impression-dominates' approach to reading than on those raters adopting other reading strategies, strongly suggests that it is not a reading strategy well suited to competency-based assessment.

Influences on rater judgment of writing ability

The finding that textual features relating to organisation and coherence were commented on by raters more than any other textual feature, particularly those adopting 'the first-impression dominates approach', has considerable support

in the literature on holistic first language writing assessment. Diederich et al (1961), Jones (1978), Freedman (1979) and Huot (1990) have all reported that when rating holistically, raters are generally most concerned with organisation. The finding that grammar was a feature most commented on by raters was not supported by the literature on first language writing assessment. In fact, past research in this area has generally concurred that raters are not particularly sensitive to grammatical features when rating holistically. However, in the field of second language writing assessment, past research has shown that, in addition to coherence and organisation, holistic raters are also influenced by grammar. Vaughan (1991), for instance, found that content, organisation and grammar were the textual features most commented on by raters while O'Loughlin (1994) found that in assessing ESL texts, ESL teachers focused primarily on content, organisation, grammar and cohesion.

It would appear, therefore, that irrespective of the assessment procedure being used (ie holistic versus competency-based), raters are focusing on similar textual features during the process of reading for assessment. The key difference, however, is that holistic raters base their assessment decisions primarily on these features, whereas competency-based raters, as the results of this study indicate, are predominantly focused on the features specified in the performance criteria. References to extraneous textual features by the raters in this study were more likely to be a function of the competency descriptor's evidence guide rather than factors directly influencing assessment decisions.

While raters adopting the 'first-impression-dominates' approach to reading commented on more textual features than other raters, analysis of their think-aloud verbal reports suggested that such comments were observational rather than judgmental. This finding is supported by Cumming (1990) who also found that raters use both interpretive and judgmental strategies in order to evaluate writing ability. Unfortunately, it was beyond the scope of this study to obtain any definitive data on the influence of these extraneous textual features on raters' assessment decisions. However, analysis of raters' think-aloud verbal reports strongly suggests that the influence of such features on raters' assessment decisions was negligible, irrespective of the reading strategy adopted.

The factor which appeared to have the most influence on the assessment decisions of the two 'first-impression-dominates' raters was not related to any extraneous textual feature at all. Rather, it was related to the issue of the promotion of learners to higher certificate levels based on the assessment of one writing sample. In this sense the 'first-impression-dominates' raters tended to

be more 'global' in terms of their assessment judgments, possibly because they were aware that their decisions had implications beyond the classroom. A possible explanation for Bill and Anne's tendency to rate globally can be found in their individual profiles. Unlike the other raters, Bill and Anne both experienced the transition from the needs-based learner-centred curriculum of the 1980s to the competency-based model of the 1990s. In terms of assessment, this meant a transition from a curriculum framework in which learning outcomes were not systematically measured or reported to one in which formal summative assessment requirements were mandated. The possibility thus exists that Bill and Anne may have brought to the assessment process a set of beliefs and attitudes about the role and purpose of second language assessment which were at odds with some of the explicit competency-based assessment principles of the CSWE. In other words, the decision-making behaviours of Bill and Anne may reflect the tension identified by Brindley (1998:47) between the need to carry out detailed individual assessments for the purposes of diagnosis and feedback to learners and the requirement to report on learners' progress against national competency standards in order to meet accountability requirements.

Assessment is a political issue, particularly competency-based assessment with its origins in the discourse of economic rationalism and its emphasis on reporting and accountability. This emphasis, combined with a trend towards tying language and literacy education funding to bench-marked competency outcome statements (see Ross, Chapter 6, this volume), has meant that teachers' assessment decisions and judgments are becoming increasingly political in nature. For some teachers, the politicisation of assessment may be an issue, particularly for those who experienced the shift from proficiency-based to competency-based models of assessment (Bottomley et al 1994). We may speculate that the possibility also exists that teachers may have some residual resistance to the changed role and purpose of assessment which may be reflected in their decision-making behaviours. Although it was beyond the scope of this study to investigate the issue of the 'resistant rater', this is an area of interest in which further research could be directed.

Conclusions and implications

Conclusions

The aim of this exploratory, qualitative study was to examine the degree to which the criteria used to directly assess written competency within the Certificate in Spoken and Written English II can ensure acceptable levels of

rater consistency and to describe and investigate the decision-making strategies and behaviours adopted by raters when assessing competency-based second language writing ability.

Given the relatively small number of raters and texts used, the results and conclusions presented in this study are naturally tentative and are not generalisable beyond the sample from which the data was drawn. In addition, in terms of the method of data collection employed by this study, it is important to note that a think-aloud verbal report is not a complete record of one's cognitive activity and thus it cannot be claimed with certainty that it represents an accurate portrayal of the thought processes used by raters when making assessment decisions. However, despite these methodological limitations, a number of tentative conclusions can be drawn which may shed light on the previously neglected field of rater consistency and judgment in the field of competency-based second language writing assessment.

Considering the question of whether the criteria used to assess written competency in the CSWE can ensure acceptable levels of rater consistency, the results of this study suggest that the overall level of rater agreement was quite high at the level of the whole text. However, it was considerably lower at the level of the individual performance criteria. Analysis of the data revealed that while raters interpreted and applied the performance criteria statements in a range of ways, they nonetheless based their assessment decisions on the extent to which the texts demonstrated the performance criteria. The finding that these raters did assess according to the assessment guidelines, while not supported in the literature, strongly suggests that raters operating within competency-based assessment frameworks do adhere to explicit statements regarding required levels of performance or competency.

In terms of the decision-making strategies and behaviours employed by raters when assessing competency-based second language writing ability, it was found that raters displayed highly individual approaches to the process of reading for the purposes of assessment. This finding has considerable empirical support in the literature. Despite the identification of three distinct reading strategies, irrespective of the reading strategy adopted, raters continued to base their assessment decisions on the performance criteria provided. However, it was noted that the performance criteria statements appeared to have a less controlling effect on raters adopting the 'first-impression-dominates' approach to reading than on raters adopting other reading strategies. In addition, raters adopting

the 'first-impression-dominates' approach to reading commented on a greater number of textual features extraneous to the performance criteria than did other raters. Moreover, they appeared to be more conscious of the possible wider implications of their assessment judgments. It was suggested that this may, in part, be related to raters' perceptions, beliefs and attitudes about the changing role and purpose of second language assessment as represented by the competency-based model. Finally, while all raters commented on a range of textual features extraneous to the performance criteria, such comments tended to be a function of observational or interpretive assessment strategies rather than judgmental ones. Again this finding was supported in the literature.

Implications

Based on these findings and tentative conclusions, a number of implications may be drawn relating to the nature of the CSWE performance criteria statements, the professional development and training of raters, and the need for further research.

In terms of the interpretation and application of performance criteria, analysis of raters' think-aloud verbal reports strongly suggested that raters had difficulty interpreting and applying some of the relativistic terminology used to describe performance. One course of action that could be considered to address this problem is to remove such terms from the performance criteria statements altogether. Analysis of the data also suggested that some prescriptive examples used in the performance criteria statements were difficult to apply. These exemplars could also be removed.

The finding that raters adopted a range of reading strategies has significant implications for the training of raters for it implies that some reading strategies may be less suited to competency-based writing assessment than others. For example, the finding that the performance criteria statements appeared to have a diminished controlling effect on raters adopting the 'first-impression-dominates' approach to reading suggests that this strategy, while resulting in a high degree of rater agreement, is not consistent with the principles of competency-based assessment. It was beyond the scope of this study to explore this issue in greater detail. However, it is clearly an area in which future research effort needs to be undertaken. If it can be established that some reading strategies result in consistently higher levels of rater agreement than others while still reflecting competency-based assessment principles, then perhaps professional

development programs could be directed towards the training of raters in the use of such approaches to reading for the purpose of assessment.

The question of the extent to which teachers' beliefs and attitudes about the role and purpose of assessment are reflected in their decision-making behaviours is one worth pursuing, particularly given the increasingly political nature of the competency-based assessment process. Future research in this area, utilising a range of data collection techniques, may provide valuable insights into this previously neglected area of research.

Finally, in order to increase our knowledge and understanding of rater judgments in the direct assessment of competency-based second language writing ability, more qualitative studies of rating processes such as the present study should be conducted. Such studies will need to investigate, on the one hand, the question of how and why raters make the assessment judgments they do and, on the other, the function and role that rating procedures, reading strategies and teacher beliefs have in this process. This is the challenge for future research.

Note

- 1 For the purposes of confidentiality, pseudonyms were used to identify the centre and the informants.

References

- Alderson, J C 1991. Bands and scores. In J C Alderson and B North (eds). *Language testing in the 1990s*. London: Macmillan, 71–86
- Bottomley, Y, J Dalton and C Corbel 1994. *From proficiency to competencies: A collaborative approach to curriculum innovation*. Sydney: National Centre for English Language Teaching and Research, Macquarie University
- Brindley, G Forthcoming. Consistency of text classification by CSWE assessors. To appear in G Brindley (ed). *Studies in immigrant English language assessment*, vol 2. Sydney: National Centre for English Language Teaching and Research, Macquarie University
- Brindley, G 1991. 'Defining language ability: The criteria for criteria'. In S Anivan (ed). *Current developments in language testing*. Singapore: SEAMO

- Brindley, G 1994. 'Competency-based assessment in second language programs: Some issues and questions'. *Prospect*, 9, 1: 41–55
- Brindley, G 1998. 'Outcomes based assessment and reporting in language learning programmes: A review of the issues'. *Language Testing*, 15, 1: 45–85
- Burrows, C 1994. 'Testing, testing, 1, 2, 3: An investigation of the reliability of the assessment guidelines for the Certificate in Spoken and Written English'. *Making Connections, 1994 ACTA-WATESOL National Conference*, 11–17
- Caulley, D, J Orton and L Clayton 1988. 'Evaluation of English oral CAT'. Melbourne: La Trobe University
- Cumming, A 1990. 'Expertise in evaluating second language compositions'. *Language Testing*, 7, 1: 31–51
- Diederich, P B, J W French and S T Carlton 1961. *Factors in the judgment of writing quality*. Princeton: Educational Testing Service
- Ericsson, K A and H A Simon 1980. 'Verbal reports as data'. *Psychological Review*, 87, 3: 215–48
- Freedman, S W 1979. 'How characteristics of student essays influence teachers' evaluation'. *Journal of Educational Psychology*, 71: 328–38
- Gipps, C V 1994. *Beyond testing: Towards a theory of educational assessment*. London: Falmer Press
- Hamp-Lyons, L (ed) 1991. *Assessing second language writing in academic contexts*. Norwood, NJ: Ablex Publishing Corporation
- Hamp-Lyons, L 1990. Second language writing: Assessment issues. In B Kroll (ed). *Second language writing: Research insights for the classroom*. Cambridge: Cambridge University Press
- Huot, B 1990. 'The literature of direct writing assessment: Major concerns and prevailing trends'. *Review of Educational Research*, 60, 2: 237–63
- Jones, B 1978. 'Marking of student writing by high school teachers in Virginia during 1978'. *Dissertation Abstracts International*, 38, 3911A.
- Jones, M 1994. Investigating consistency in assessment in the Certificates in Spoken and Written English. *Draft Project Report*. Sydney: NSW AMES
- Milanovic, M, N Saville and S Shuhong 1996. A study of the decision making behaviour of composition markers. In M Milanovic and N Saville (eds). *Studies in Language Testing 3: Performance testing, cognition and assessment*. Selected papers from the 15th Language Testing Research Colloquium. Cambridge: Cambridge University Press
- New South Wales Adult Migrant English Service 1998. *Certificates in Spoken and Written English, I, II, III and IV*. 2nd ed. Sydney: New South Wales Adult Migrant English Service
- Nisbett, R E and T D Wilson 1977. 'Telling more than we can know: Verbal reports on mental processes'. *Psychological Review*, 84, 3: 231–57
- O'Loughlin, K 1994. 'The assessment of writing by English and ESL teachers'. *Australian Review of Applied Linguistics*, 17, 1: 23–44
- Vaughan, C 1991. Holistic assessment: What goes on in the rater's mind? In L Hamp-Lyons (ed). *Assessing second language writing in academic contexts*. Norwood, NJ: Ablex Publishing Corporation
- Weigle, S 1994. 'Effects of training on raters of ESL compositions'. *Language Testing*, 11, 4: 197–217

6

Individual differences and learning outcomes in the Certificates in Spoken and Written English

Steven Ross

Introduction

Investigations of learning outcomes in the Certificate in Spoken and Written English (CSWE, NSW AMES 1998) competencies between 1996 and 1998 have revealed considerable variation across individual clients (Ross 1997, 1998). Comparisons of patterns of successful certificate achievement have also indicated considerable variation in the impact of the 510 hours of instruction on CSWE program outcomes. Results of the outcomes research suggest that there may be individual differences factors that, beyond the initial assessment of ASLPR level, differentially promote and even constrain the likelihood of eventual achievement of competencies and certificate awards. The present chapter aims to identify some of these factors so as to promote a process of classifying clients at placement into the AMEP (Adult Migrant English Program) which can more accurately reflect individual profiles and needs.

Individual differences in language learning

From a program administration viewpoint, it is often puzzling as to why CSWE clients, given equal instruction and starting from the same proficiency level, do not complete CSWE competencies at a constant pace. In order to understand this phenomenon, it needs to be recognised that adult language learning is not a linear process. Considerable second language research (Ellis 1996; Larson-Freeman and Long 1991) has suggested that there are cognitive stages of language learning through which adult second language learners appear to progress differentially, often independent of the influence of their first language (Pienemann 1984; Pienemann and Johnston 1986). Other accounts of this phenomenon have posited constellations of affective factors (Anderson 1982; Gardner 1985) which serve to slow down the process of

acquisition, leading to learning ‘plateaus’ — often reflected in fossilisation of interlingual forms such as idiosyncratic grammars (Gass and Selinker 1994). Arrested language learning indicates that there may be deeper social and psychological factors that influence the speed and ultimate attainment of adult second language learners (Schumann 1978). Factors such as language learning motivation and aptitude (Carroll and Sapon 1959; Erhman 1996; Skehan 1991; Sasaki 1996) have been the object of a good deal of research. Further, there are individual difference factors related to experience prior to exposure to formal instruction in adulthood (Bialystok 1997; Birdsong 1992; Singleton 1989; Long 1990; Scovel 1988). Experiential factors suggest a ‘readiness’ phenomenon that is likely to interact with affective, social, and cognitive influences (Stern 1976; Walberg 1978).

The complex constellation of factors described above is not well understood, however, and tends to be subsumed under the general rubric of ‘individual differences’ (Bley-Vroman 1989; Skehan 1989, 1991). Whether case studies of language learning are involved (Schumann 1978; Schmidt 1983), or cross-sectional differences in language learning outcomes are examined, the impact of individual differences appears to be pervasive. Individual differences are thus the starting point for the analysis of differential outcome patterns in the CSWE. Since this analysis is essentially *ex post facto* — relying on archival data — the focus will of necessity be on experiential sources of individual differences, in other words, on biographical variables. A more comprehensive picture of the phenomenon would need to include an examination of a much wider range of affective, social-psychological, cognitive and experiential factors (Gardner and Tremblay 1994).

Several variables related to individual differences were considered in the analyses. All of these variables were derived from the AMEP Research Management System (ARMS) database. Those considered in this cross-sectional research on individual difference factors are sex, language distance, age, length of residence in Australia, education in the home country, tertiary education experience in the home country, and hours of instruction in the CSWE program. The aim of the research is to identify the most important factors that may yield optimal information for predicting differential outcomes in the CSWE program.

Variables

The following is an outline of the procedures used in the CSWE profiling project, which is intended to provide an optimal set of diagnostic variables

influencing differential rates of achievement on the CSWE. The variables of interest are either part of the ARMS database, or derived from it. These variables are also known in second language research as being relevant to processes impacting on outcome differences in second language acquisition by adults.

The ARMS database was edited so as to yield samples with complete records in each CSWE level. The key criterion for inclusion into the profile database was non-missing date of arrival record in the ‘date of arrival’ (DOA) field in ARMS. Thereafter, based on the other fields in ARMS, a few new categories were generated for the profile project. These were:

- **Language Distance** (LDist), which was operationalised as an ordinal-scaled variable. Criteria for the scaling were:
 - a) orthography (alphabetic, syllabic, ideograph);
 - b) canonical word order (SVO, SOV, OSV, etc.);
 - c) typological grouping (Germanic, Latinate, Slavic, Altaic, Sino-Tibetan, etc).

Languages closest to English were given the highest proximity value (eg Dutch or German). The most distant languages (eg Cantonese and Korean) were given the lowest.

- **Tertiary Education** (Tert) was updated from ARMS so as to provide a parallel variable to years of education (Edu) already contained in ARMS.
- **Length of Residence** (LOR) was calculated from the reported date of arrival in Australia. The scale is relative to the *most recent* arrival in the sample taken. Thus, if the most recent arrival in a given sample is ‘01/12/97’, all other clients in the sample are scaled to show a LOR relative to this date. The unit of measurement is ‘days in country’.
- **Competencies Achieved** (Comps) were tallies taken directly from ARMS. However, for the purpose of consistency checking, even and odd competencies were tallied to create both a total count and an odd plus even tally for some analyses and reliability checking.

The initial data set retains all of the ARMS information plus the new categories: Client ID, Language Code, Language Descriptor, Sex, Age, Migration Category, Country of Birth, Education in the Home Country, Randomisation Code, Hours in Program, Awards, Level & Band, CSWE Competencies, Certification Code. For the baseline data set in each certificate, a single master

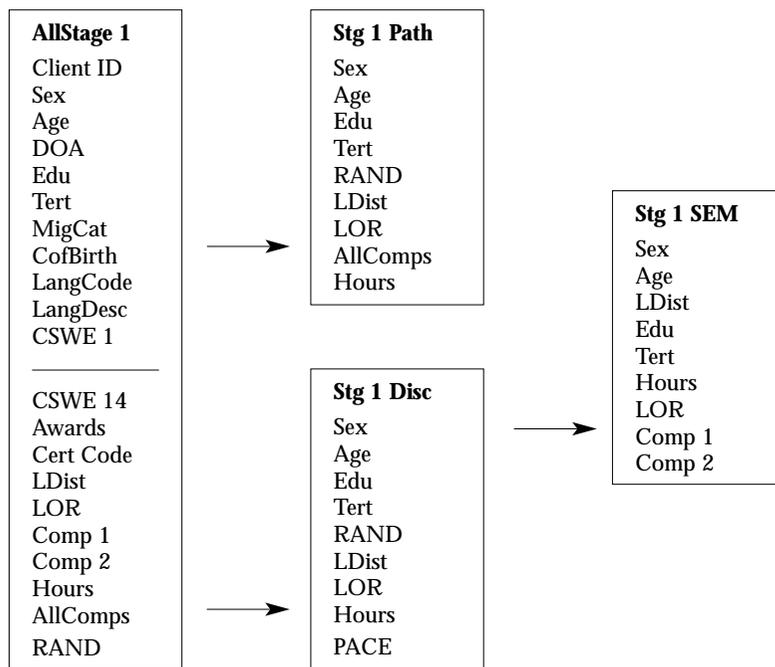


Figure 5: ARMS data categories

file was created. Sub-files were then created for each type of analysis. The derivation of sub-files is shown in Figure 5.

Databases

The three sub-files were used for different types of analyses; **Stg1Path** was the input to a recursive (time-mediated) path analysis. The goal of the path analysis is to identify direct and indirect paths (standardised regression weights) to the total number of competencies achieved. The RAND code serves to provide a basis for cross-validation through parallel analyses of odd and even case records. As in outcomes descriptions, the odd and even tallies provide an index of replicability.

Stg1Disc is the input file to a three-group linear discriminant function analysis. The goal here is to estimate discriminant function weights for each direct discriminator of *slow*, *average* and *fast* paced learner categories. The variable

PACE is a code indicating the total number of CSWE competencies achieved — though here the coding variable corresponds to fewest (slow), median (average), and most (fast). Again, RAND serves as the basis of cross-validation and model testing. In the test set (eg EVEN), discriminators of PACE are forced into the discriminant model after hours has been used to define the PACE variable (described in detail below). Significant predictors of the PACE categories are then recombined into an optimal discriminant expression. In the cross-validation run, the weights from the test set are used to predict membership in the observed PACE categories in the ODD data set.

The **Stg1SEM** data set includes nine measured variables which are the basis of input into a structural equation model. The model is designed to isolate major groupings of ARMS variables into subsets of ‘indicators’. These are then tested for fit to the observed data. The nine variables are organised into four measurement models with measured indicators and latent variables. Each of these compete as influences on the achievement of CSWE competencies so as to test the stability of differences related to characteristics of the individual (LDist and Age) relative to home-country education (indicated by Edu and Tert). The same structural model is then compared across the three CSWE levels while holding LOR and Hours in Program constant.

A further collation of data was performed in order to provide a back-up set of analyses with non-linear statistical assumptions. This was accomplished with logistic regression. Here, the CSWE tallies of competencies were recoded to indicate ten or more competencies achieved. This code represented CSWE *certification*. A sum of less than ten competencies was coded as a *non-certification*. Thereafter, all of the co-varying factors were used as predictors of the dichotomous certification outcome.

The data were also collated into distinct subsets based on different native languages. Here, Arabic, Bosnian, Mandarin, and Vietnamese were sampled ($n = 1200$ each) with a view to assessing the cross-linguistic stability of the putative individual difference factors. The main aim of the non-linear analysis was to examine whether the same factors promote or constrain the likelihood of certification across first language groups. The cross-language analysis allows for a close-up examination of the predictors and the extent to which they may differentially function across first language groups.

The logistic regression analysis was also done for each migration category. Here, the five categories of migration were collated into separate data sets for

the purpose of exploring whether the same factors promote or constrain each group's probability of CSWE certification. We may assume that if predictors are stable across language groups and migration categories, they will serve as the most important key variables in the deeming process without particular qualification.

Analyses

The preliminary analysis was devised in order to obtain a general picture of factors that have direct and indirect impacts on the acquisition of CSWE outcomes. The first method used, a recursive path analysis, orders the measurements in a chronological sequence so as to find individual differences in each time co-varying with consequential differences. These are then used to estimate relative influences on CSWE in direct and indirect paths. This analysis can be considered exploratory, since it is restricted to Certificate I competencies.

The first wave of factors are sex, age, and language distance, which are considered immutable aspects of each client's background. They also represent factors that exist prior to migration to Australia. The second wave is also made up of pre-migration factors, but can be understood as those potentially affected by some or all of the first wave of factors. The second wave of factors, education in the home country and years of tertiary education, may also have a potential 'downstream' impact on program participation, and may be influenced by sex, age, or language distance.

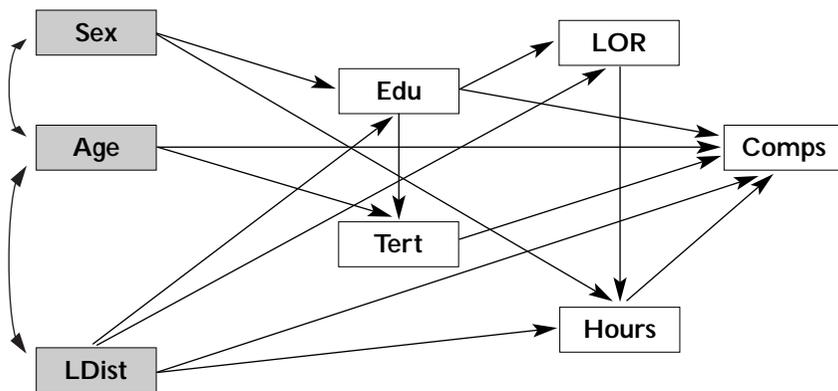


Figure 6: CSWE Certificate I recursive path model of competencies achieved

Figure 6 shows the final path model used. The single-headed one-direction arrows indicate a non-random path ($p < .05$) from an earlier individual difference factor to a subsequent factor. The rightmost variable is the total number of CSWE competencies achieved. Paths may be direct from a variable to CSWE competencies, or indirect, if they are mediated by an intervening variable. Curved paths indicate correlations between pre-migration factors.

The recursive path analyses were done twice for a randomly sampled Certificate I data set ($n = 6948$). This data set was split into odd and even halves for the two separate analyses. The object of the split of tallies was to find consistent path (standardised regression weights) across the samples. These are reported in Table 33.

Table 33: CSWE Certificate I cross-validation results

Correlations	Even (N = 3488)	Odd (N = 3460)
Sex with Age	-.10	-.08
Age with LDist	-.04	-.015
Indirect Paths		
Sex to Edu	-.041	-.05
LDist to Edu	.085	.05
Age to Tert	.226	.223
Edu to Tert	.229	.250
LDist to LOR	-.135	-.100
Edu to LOR	-.051	-.070
Tert to LOR	-.055	-.070
Sex to Hours	.089	.070
Edu to Hours	.048	.050
LDist to Hours	.056	.090
LOR to Hours	.160	.160
Direct Paths		
Age to Comps	-.174	-.164
LDist to Comps	.054	.034
Edu to Comps	.135	.136
Tert to Comps	.097	.072
Hours to Comps	.679	.685
Regressions	$R^2 = .538$	$R^2 = .537$

It can be seen from Table 33 that the largest indirect paths are between age and education, and between education and tertiary education. The former indicates that the older the migrants are in the sample, the more likely they are to have completed relatively more years of education in their home countries. The path from education to tertiary education suggests that clients with the most years of basic education are more likely to have relatively more tertiary education. Language distance shows a negative path to length of residence. This suggests that recent migrants in the CSWE program tend to be native speakers of languages relatively distant from English. Length of residence in Australia has a positive path to hours of tuition in the program. This comes as no surprise.

Certainly not all factors have a direct influence on CSWE outcomes (Comps). Indirect paths can be identified by connecting single-headed arrows from downstream variables which have an impact on CSWE. If downstream variables have arrows pointing at them, we can infer that the leftmost variables with paths connected downstream affect CSWE competency achievements indirectly. By way of example, we can observe this phenomenon in the path from Sex to Education (small and negative), which suggests that female migrants have slightly less basic education than do males. This factor does not have a direct influence on CSWE. Rather, we find that education has a direct influence (.135) on CSWE competencies.

Direct paths, in contrast, show unmediated influences on CSWE outcomes. Here we can see that age has a direct negative impact on CSWE. Older learners progress at a slower rate of achievement than younger learners. Education in the home country has a direct positive influence on the achievement of competencies. Education may indicate first language literacy, which may be a powerful prerequisite for adult second language reading development. It also may be an indicator of cognitive development for adult learners.

The most obviously important factor is tuition hours in the CSWE program. Given the nature of the competency-based curriculum, hours of tuition lead directly to CSWE competencies in such a manner that it is not possible to assess achievements of non-participants in the CSWE program. The path between Hours and Comps can also be used as a 'control' so as to allow other individual factors to 'compete' while holding Hours constant. The structural equation section of this study is conducted in this manner.

Structural equations

Following the recursive path analysis, the eight variables of interest derived from ARMS were then recombined into measurement models (measured indicators of latent variables). The combinations of measured variables and latent variables were then combined into a structural equation model. Structural equation models facilitate empirical analyses of competing explanatory influences on variables of interest. In the CSWE data, the acquisition of competencies are the main outcomes of interest. Previous educational experiences in the home country (Education and Tert) are compared with other potentially explanatory variables. Exposure to English, as indicated by length of residence (LOR) and hours of instruction in the AMEP (Hours) are modeled simultaneously with previous education, age, and language distance in order to discover the most important influences on the acquisition of competencies. In these analyses, we consider length of residence and hours of CSWE tuition to be 'given' influences on CSWE competency achievement. We will therefore hold this measurement model labeled Experience (Exper) constant for the purpose of comparing the three other paths thought to influence CSWE outcomes.

Certificate I competencies

Variables from the ARMS database were combined into sets of indicators. The indicators (rectangles) together may create a latent factor (oval). For instance, hours of instruction in the AMEP and length of residence in Australia together indicate a factor which we will call 'exposure' to English. Home country education and tertiary education together indicate another factor called 'learning'. Client age and native language are singleton variables which cannot be combined into an indicator set, so these remain as measured influences (rectangles) on the acquisition of competencies. Note also that CSWE is here defined as a latent variable indicated by the odd and even competencies from the ARMS database. For Certificate I, individual difference data from 9447 learners were randomly sampled from ARMS.

The model (shown as Figure 7) allows us to test the relative influence of single variables and measured indicator sets on the acquisition of CSWE competencies. The indicators and variables create competing paths of influence on CSWE. Each path from an indicator or variable is unique in that it provides a unique and standardised index of that indicator or variable's influence on CSWE while all other variables and indicators are held constant. The figure on each path shows the magnitude of the indicator or variable's influence in standard deviation units of influence on CSWE.

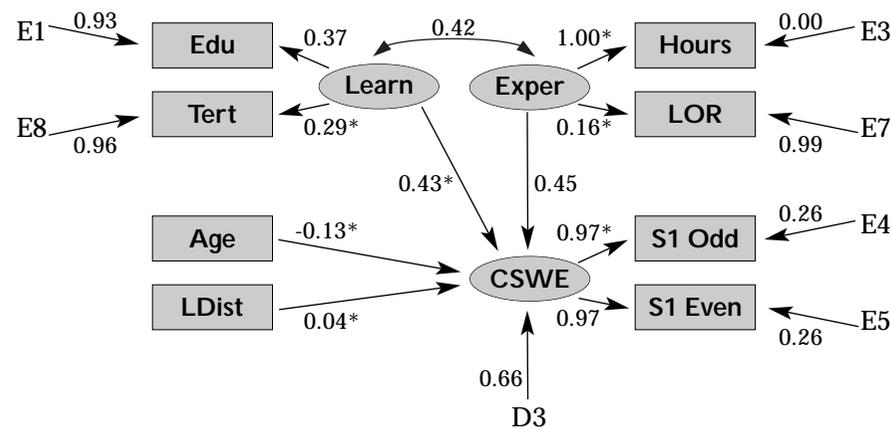


Figure 7: Structural Equation Model: Certificate I

A double-headed curved arrow between latent indicators (ovals) indicates that there is a non-causal correlation between the indicator sets.

The structural equation model for Certificate I indicates that learning experiences prior to migration to Australia (Learn — indicated by years of basic education and years of tertiary education) is the most important factor when exposure to English (Exper — indicated by hours of CSWE instruction and length of residence in Australia) is controlled. Client age is a constraint on CSWE outcomes, while language distance (LDist) appears to have little effect. Each path is shown as a standardised regression coefficient, which indicates the percentage change in one standard deviation unit of the CSWE that increases or decreases with one standard deviation change in the latent or measured variable from which the arrow comes. The double-headed curved arrow shows the correlation between pre-migration learning and program participation. This figure suggests that better educated clients are likely to participate in the CSWE program longer than less educated clients.

The cognitive development variable suggests that the tasks of literacy and adult second language acquisition are promoted by pre-migration literacy in the first language. Age, in contrast, can be seen to slightly constrain the achievement of CSWE competencies. The ordinal code for language distance, in contrast appears to have relatively little unique effect on CSWE when other paths are controlled.

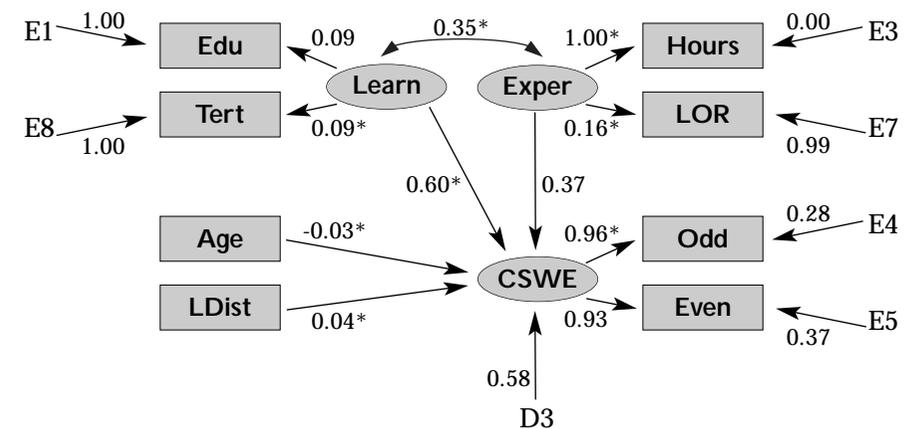


Figure 8: Structural Equation Model: Certificate II

Certificate II competencies

In order to test the consistency of the same model at the next CSWE level, new data were modelled using the same hypothesis as that shown in the Certificate I structural equation. Here, 7555 Certificate II learners were randomly sampled from the ARMS database. Figure 8 shows the results of the analysis on Certificate II.

The paths appear to differ in magnitude in Certificate II. The relative importance of the three competing paths, however, is consonant with Certificate I. There is an increased importance of pre-migration learning in Certificate II. Age appears to be less of a constraint at this Certificate level, while language distance remains insignificant as an influence on CSWE outcomes.

Certificate III competencies

Certificate III tallies are based on the total number of competencies achieved in any of the three streams of the Certificate III syllabus — Community Access, Further Study, or Vocational English. Since clients are placed into Certificate III after the Australian Second Language Proficiency Ratings (ASLPR) assessment, or have progressed through the curriculum to reach the highest level, it is likely that Certificate III clients are faster paced learners than clients in Certificates I or II.

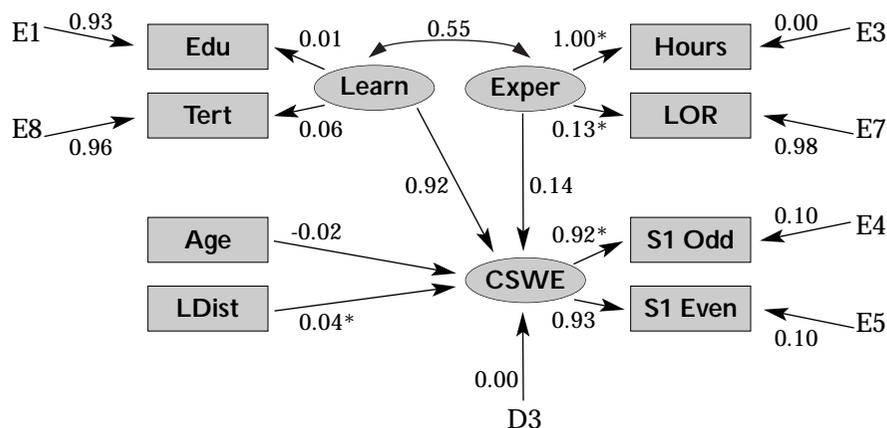


Figure 9: Structural equation Model: Certificate III

In order to test the hypothesised influences on CSWE outcomes, we once again use the model devised for Certificates I and II. The result of the same structural equation is observed in Figure 9.

The pattern observed in Certificates I and II is again seen in Certificate III ($n = 4417$). It appears that learning in the home country again correlates with participation in the CSWE program. Pre-migration learning here, moreover, has by far the largest impact on CSWE. Age and LDist appear to be insignificant in the highest level of the CSWE curriculum.

The structural equation model tested suggests that age constrains CSWE in Certificate I, and pre-migration learning serves to promote it. The influence of age diminishes in Certificates II and III, while pre-migration learning becomes increasingly important. The goodness-of-fit statistics shown in Table 34 indicate that the most stable structural equation model is that constructed for Certificate I. Certificates II and III are less well fitting, and are therefore subject to revision if paths are respecified.

Table 34: Structural Equation Model outcomes Certificates I – III

Certificate	Comparative Fit Index	Chi-Square Probability
1	.957	<.001
2	.855	<.001
3	.844	<.001

Discriminant analysis

A third linear analysis was undertaken with the goal of finding the optimal predictive weights associated with individual difference variables gleaned from the ARMS database. Here again, Certificate I data was used for the purpose of model building. The linear discriminant analysis aims to identify measured variables that best discriminate among distinct classifications. In this study, the classifications of interest are empirically-derived profiles of 'slow', 'average', and 'fast' CSWE Certificate I clients. The classification method was based on identifying records in the ARMS database that could be distinctly classified according to the following criteria:

Slow Pace

CSWE clients who had achieved a number of competencies which was *fewer* than one standard deviation below average Certificate I competencies after *more* than a standard deviation of hours of tuition above the Certificate I average.

Average Pace

CSWE clients with average numbers of competencies achieved — defined as within one-half a standard deviation from the Certificate I average, and who had used within one-half a standard deviation from the Certificate I average of tuition hours.

Fast Pace

CSWE clients who had used *less* than one standard deviation of tuition hours in Certificate I, but who had achieved *more* than one standard deviation above the Certificate I average number of competencies. Figure 10 sketches the manner in which the PACE coding was devised.

The Certificate I data set was split into two equal halves based on the last digit of the client ID number. The even-numbered half was used as the training set, and the odd-numbered half was saved as the test set. In both data sets a new variable PACE was included so as to make the comparison of training set results relative to the test set results. The PACE variable was operationalised as the code denoting each client's empirical learning pace as evidenced through competencies achieved relative to hours of CSWE tuition used. The PACE criteria are shown in Figure 10.

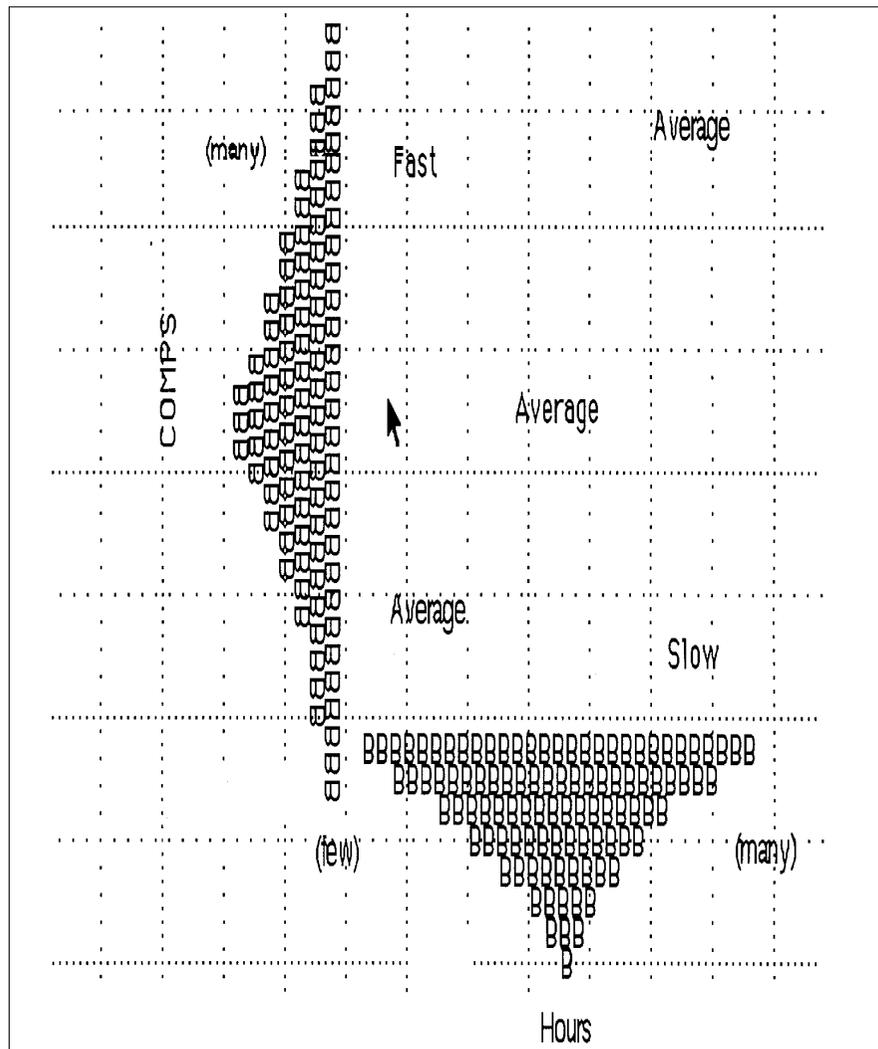


Figure 10: PACE criteria

For the purpose of optimal classification, equal numbers of PACE code cases were retained in the odd and even halves of the data set. Both the odd and even halves of the Certificate I data set used for this phase of the analysis had

approximately 1000 cases each with roughly equivalent numbers of slow, average, and fast paced learners.

A linear discriminant analysis was first performed on the even half of the Certificate I data set. This analysis included all of the variables shown in Figure 5 in the box labelled Stg1Disc. Not all of the ARMS variables, however, appear to provide discriminating information about individual differences in the PACE variable. The criterion for retention in the model was set at $p < .01$, which allows a reduction of variables in the training set for the final discriminant model. Table 35 summarises the most significant variables.

Table 35: Discriminant function analysis summary

DISCRIM No. of vars. in training set model: 4; Grouping: PACE (3 grps)
 STATS Wilks' Lambda: .86881 approx. F (8,1976)=17.993 $p < 0.0000$

N = 994	Wilks' Lambda	Partial Lambda	F-remove (2,988)	p-level
Edu	.9068388	.9580685	21.62071	.0000000
Age	.9091569	.9556258	22.93871	.0000000
LOR	.8852504	.9814329	9.34569	.0000953
Tert	.8910021	.9750974	12.61605	.0000039

It appears that the main variables showing an influence on Comps in the first two linear analyses done, the Path Analysis and the Structural Equation Model of Certificate I, also serve to discriminate among the learning pace classifications. This result constitutes corroborating evidence that a subset of ARMS variables serve to differentiate among CSWE outcomes. The extent of the classification accuracy in the discriminant analysis can be observed graphically. Figure 11 shows a scatterplot of the individual learning pace group members relative to their own group centroids. The latent roots are the average discriminant functions for each group on each function.

Figure 11 provides a visual display of the extent of discrimination among the empirically-derived learning PACE variable. As the first step in the discriminant analysis, the training study indicates that the subset of four variables (home country education, age, length of residence in Australia, and tertiary education) are the best discriminators of the actual acquisition of CSWE. The next step of the analysis was to examine the accuracy of the training set classifications on the test set, which was in fact the odd halves of the Certificate I sample.

Root 1 versus Root 2

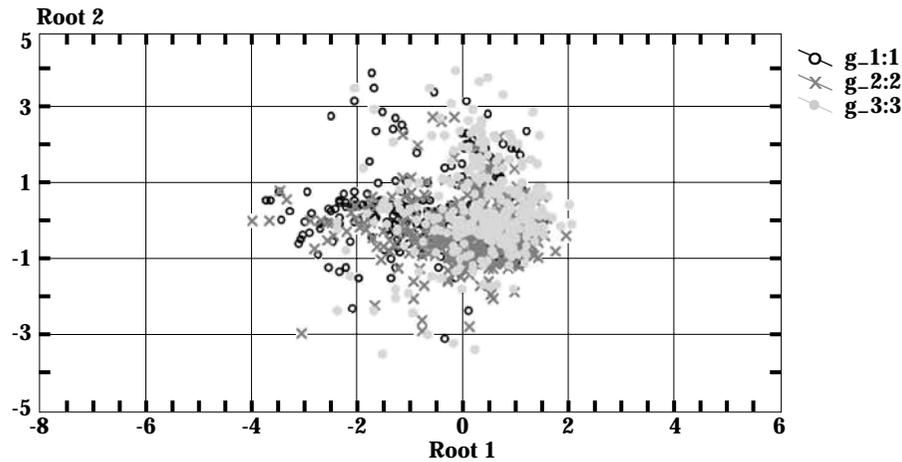


Figure 11: Slow (1), Average (2) and Fast (3) Group Scatterplot

Each client in the test set (odd half) was classified with a *slow*, *average* or *fast* learner profile based on the method shown in Figure 10. In this analysis, the discriminant function weights from the four variables with significant discriminant power were cross-multiplied with the actual ARMS data for each of the clients in the odd halves of the Certificate I database. For instance, each client's home country education in ARMS was converted into a new variable 'eduwght' which had three possible 'weights' based on the training set outcome:

```

replace all   eduwght1   with   edu*.7474
replace all   agewght1   with   age*.2644
replace all   lorwght1   with   lor*.0084
replace all   terwght1   with   tert*.1912
replace all   eduwght2   with   edu*.8022
replace all   agewght2   with   age*.2463
replace all   lorwght2   with   lor*.0105
replace all   terwght2   with   tert*(-.0788)
replace all   eduwght3   with   edu*.8563
replace all   agewght3   with   age*.2034
replace all   lorwght3   with   lor*.0070
replace all   terwght3   with   tert*.0300
    
```

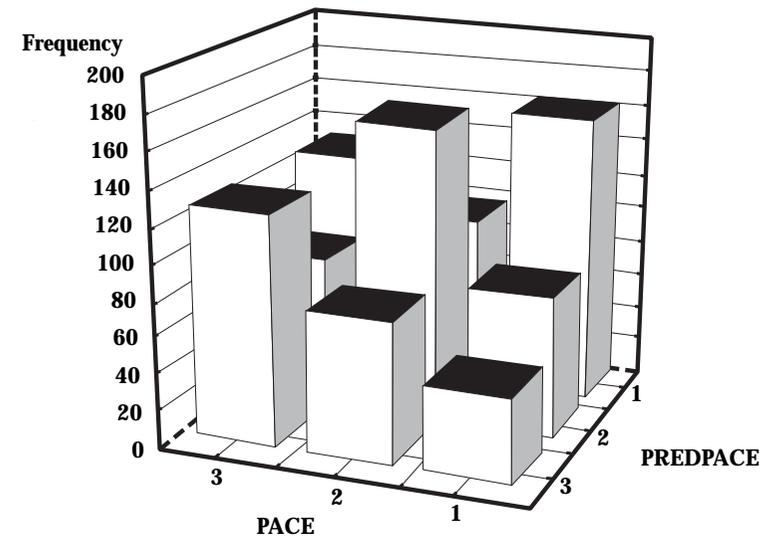


Figure 12: Histogram of PACE predictions for the test set (odd half)

After the cross-multiplication based on the discriminant weights from the training set, each client's classification 'weight' was derived by adding the constant for each PACE category to the sum of the client's data multiplied by the classification weights from the training study:

```

replace all   awght   with   acon+eduwght1+agewght1+lorwght1+terwght1
replace all   bwght   with   bcon+eduwght2+agewght2+lorwght2+terwght2
replace all   cwght   with   ccon+eduwght3+agewght3+lorwght3+terwght3
    
```

The final step in the preparation of the test set was to predict the learning pace of the odd half sample based on the discriminant functions derived from the even half. This was done with the creation of a new variable 'predpace', or prediction of learning pace. A client would be predicted to be slow, average or fast pace if his/her individual difference profile data in ARMS had a 'weight' characteristic of a slow, average or fast (a, b or c) profile client:

```

replace all   predpace   with 1   for awght > bwght .and. awght > cwght
replace all   predpace   with 3   for cwght > bwght .and. cwght > awght
replace all   predpace   with 2   for bwght > awght .and. bwght > cwght
    
```

The empirical test of the predicted pace was conducted by cross-tabulating the actual learning pace in the test set (odd half) by the predicted learning pace derived from the training set (even half). Figure 12 shows the ‘hit rate’ or accuracy of the predicted versus observed PACE outcomes for the test set.

The highest bars coinciding with the classifications reflect prediction accuracy. It can be observed that the coincidence of predicted (*PREDPACE*) and observed (*PACE*) is highest for the slow (‘1’) and average paced (‘2’) clients in the test set. The fast profile (‘3’) has the largest error rate. This outcome suggests that there is some confound in the prediction process for the classification of fast paced learners in Certificate I of the CSWE program.

The three linear methods used to explore individual differences (path analysis, structural equation modelling, and discriminant analysis) all have distributional assumptions that cannot be consistently met across the three levels of the CSWE curriculum. For this reason, we will consider also the same variables using non-linear regression, which does not make strong assumptions about the distribution characteristics of the predictors. The non-linear models will thus serve to cross-validate the pictures of the relative importance of individual difference factors we have observed so far.

Non-linear models

The first set of logistic regressions was performed on four different language groups. These particular groups were chosen in order to capture the language distance factor ranging from relative proximity to English (Bosnian) to distant from English (Mandarin). The non-linear analyses were done on Certificates I and II only because there are insufficient cases in the database for accurately modelling Certificate III with its three different tracks.

The logistic regression functions analogously to linear regression in that the predictors can be evaluated as independent sources of influence on the outcome. Here, the outcome is the likelihood of certificate achievement in the respective CSWE levels. The statistic of interest is the odds ratio, which indicates the odds of a change in the probability of certification if there is a one unit change in the predictor variable. For instance, with an *odds ratio* of 1.5 for Tert, there would be a positive change in the odds of certification by a multiplicative factor of 1.5 for each year of tertiary education.

The convention we will follow in these analyses is to mark the statistically

significant odd ratios with a ‘+’ in order to indicate that there is a positive influence on certification, and a ‘-’ to indicate that there is a significant constraint on certification with one unit change in the predictor variable.

The first analysis is Certificate I certification for the four language groups sampled from ARMS. The positive influences are Edu and Hours of instruction (Table 36). Since hours is a ‘given’ influence on certification, it will not be discussed in subsequent analyses. Age is a consistent constraint on certification at this level. As age increases by one year, there is a monotonic change in the likelihood of *not* achieving Certificate I. It is worthy of note that Edu is the most influential predictor for three of the language groups. For the Mandarin speakers, however, it appears that education does not influence certification as expected. This may suggest that reporting criteria among the Mandarin speakers in Certificate I may be different from other clients, or that the impact of basic education in the PRC may not influence second language development the way it does for other clients. Client Band appears not to provide consistent information about progress, although it may be noted that there is considerable difference in the relative size of the odd ratios among the groups.

Table 36: Odds ratios for Certificate I by language group

Certificate I	Arabic	Bosnian	Mandarin	Vietnamese
Sex	0.986	0.915	1.135	0.961
Edu	1.121+	1.141+	1.038	1.148+
Tert	1.150+	1.081	1.088	1.124
Age	0.972-	0.975-	0.961-	0.975-
Hours	1.009+	1.008+	1.009+	1.009+
Band	1.140	0.800	1.249	1.143
LOR	0.999-	0.999-	1.000	1.000

The picture for Certificate II is similar (Table 37). At this level we find that education (Edu) shows a consistent influence on the outcomes for all of the groups. Age once again constrains achievement. When we compare the influence of Tert, only the Bosnian group’s reported tertiary education has an impact on CSWE certification likelihood. In Certificate I, only the Arabic group’s Tert showed an effect. From these discrepancies, we can infer that tertiary education may in fact vary across groups — having predictive power for some, and no power for others.

Table 37: Odds ratios for Certificate II by language group

Certificate II	Arabic	Bosnian	Mandarin	Vietnamese
Sex	1.003	1.181	1.147	1.030
Edu	1.242+	1.295+	1.099+	1.410+
Tert	1.150	1.172+	1.085	1.154
Age	0.958-	0.930-	0.947-	0.953-
Hours	1.011+	1.014+	1.010+	1.011+
Band	1.445	1.467+	1.273	1.252
LOR	1.000	1.001	1.000	1.000

The second phase of the non-linear analysis involved comparing different migration categories in order to see if the same set of predictors across language groups turn up when migration category is the basis for data selection. Table 38 lists the outcomes of the logistic regression analysis based on migration category for Certificate I outcomes.

Education (Edu) was found to be significant for four out of the five migration categories. Among the Skilled category of migrants, education does not seem to be a predictor, possibly because there is less variance among clients in terms of basic education among these CSWE clients. Tertiary education shows an impact for Humanitarian and Skilled migrant categories, but not in the expected directions. Tertiary education appears to be a constraint among the Humanitarian class clients, while it appears to advantage the Skilled class clients. Further investigation of this discrepancy could reveal unique characteristics of the Humanitarian class clients.

Table 38: Odds ratios for Certificate I by migration category

Certificate I	Concess	Family	Humanit	Skilled	Other
Sex	1.081	1.044	1.000	1.233	1.034
Edu	1.136+	1.105+	1.145+	1.021	1.065+
Tert	1.051	1.021	0.799-	1.097+	1.120
Age	0.989	0.979-	0.994	0.986	0.987
Hours	1.008+	1.007+	1.006+	1.007+	1.008+
Band	0.923	0.892	1.145	1.322+	1.472+
LDist	0.998	0.929	1.029	0.886-	0.940
LOR	1.000	0.999-	1.000	1.000	0.999-

Another interesting outcome in this analysis is that age does not appear to be a constraint among the different client migration categories. We noted that among language groups, age appeared to be a near universal constraining factor. While the odds ratios for age across groups are less than one (meaning a small constraint), only one of the five classes shows a significant effect.

The situation for Certificate II follows the overall pattern revealing basic education to be a significant promoting factor. But for Certificate II, the age factor re-emerges as a significant constraint in all five categories. Tertiary education promotes certification for the Concessional, Skilled, and Other migrant categories. In Certificate II, Band also emerges as predictor of certification for three out of the five categories.

Table 39: Odds ratios for Certificate II by migration category

Certificate II	Concess	Family	Humanit	Skilled	Other
Sex	0.761	0.861	0.713	1.470	1.460
Edu	0.990	1.187+	1.348+	1.122+	1.250+
Tert	1.211+	0.902	1.082	1.173+	1.251+
Age	0.930-	0.967-	0.944-	0.943-	0.937-
Hours	1.010+	1.009+	1.011+	1.010+	1.011+
Band	1.633+	1.524+	1.573+	1.072	1.049
LDist	1.039	1.276+	1.090	1.169+	1.002
LOR	1.000	1.000	1.000	1.000	1.000

The logistic regressions based on language groups and migration category suggest that there are within-subgroup individual difference factors that exist at the micro level of analysis. They also reveal that home-country education is the most consistent covariate with CSWE success, while age tends to be a constraint. The similarity of the odds ratios across the language groups corroborates the results of the linear analyses — that language distance as it is constructed here does not appear to show a consistent influence on CSWE outcomes one way or another.

Summary

The linear and non-linear analyses of the CSWE competencies have tested a number of ARMS variables as useful predictors of outcomes. These findings are summarised in Table 40.

Table 40: Summary of individual differences analyses

Analysis Method				
<i>Variables</i>	<i>Path Analysis</i>	<i>Structural Equations</i>	<i>Discriminant Analysis</i>	<i>Logistic Regression</i>
Edu	✓	✓	✓	✓
Age	✓	✓	✓	✓
Tert			✓	✓
LDist		✓		
LOR		✓		
Sex				

These findings have a range of implications for the process by which AMEP clients are classified at entry to the Program (currently known as 'deeming') and for the deeming procedure itself. The relative importance of education in the home country as a consistent indicator of CSWE success cannot be understated. This variable is the most important of those in the ARMS database. Client age is also of clear importance in streaming the clients into Levels and Bands. If there is a need to simplify the deeming process to as few factors as possible, the initial intake interview (presumably ASLPR) plus accurate information about home country formal education and client age would be minimally sufficient to optimally place clients.

If a more thorough deeming process is feasible and desired, other information about tertiary education and length of residence in Australia could be added. However, whether or not the addition of these variables would make the procedure more accurate than the more abbreviated procedure is a question for future research.

Finally, the logistics of implementing a standard deeming procedure probably demand a standardised procedure. The ideal solution would be a computerised process that prompts the intake interviewer for information about new clients. This information would then be input to an 'expert system' which could then identify the optimal CSWE Level and Band for a new client. Such a system is likely to expedite the process of deeming and lead to a better delivery of the CSWE curriculum.

References

- Anderson, P L 1982. 'Self-esteem in the foreign language: A preliminary investigation'. *Foreign Language Annals*, 15: 109–14
- Bialystok, E 1997. 'The structure of age: In search of barriers to second language acquisition'. *Second Language Research*, 13: 116–37
- Birdsong, D 1992. 'Ultimate attainment in second language acquisition'. *Language*, 68: 706–55
- Bley-Vroman, R 1989. What is the logical problem of foreign language learning? In S Gass and J Schachter (eds). *Linguistic perspectives on second language learning*. Cambridge: Cambridge University Press
- Carroll, J B and S M Sapon 1959. *Modern Language Aptitude Test: MLAT*. New York: Psychological Corporation
- Ellis, R 1996. *The study of second language acquisition*. Oxford: Oxford University Press
- Erhman, M 1996. *Understanding second language learning difficulties*. Thousand Oaks, CA: Sage
- Gardner, R 1985. *The social psychology of second language learning: The role of attitude and motivation*. London: Edward Arnold
- Gardner, R C and P F Tremblay 1994. On motivation, research agendas, and theoretical frameworks. *Modern Language Journal* 78, 3: 359–368
- Gass, S and L Selinker 1994. *Second language acquisition*. Hillsdale, NJ: Lawrence Erlbaum
- Larsen-Freeman, D and M Long 1991. *An introduction to second language acquisition research*. London: Longman
- Long, M 1990. 'Maturational constraints on language development'. *Studies in Second Language Acquisition*, 12: 251–85
- New South Wales Adult Migrant English Service (NSW AMES) 1995. *Certificates in Spoken and Written English*. Sydney: New South Wales Adult Migrant English Service
- New South Wales Adult Migrant English Service 1998. *Certificates in Spoken and Written English, I, II, III and IV*. 2nd ed. Sydney: New South Wales Adult Migrant English Service

- Pienemann, M 1984. 'Psychological constraints on the teachability of languages'. *Studies in Second Language Acquisition*, 6: 186–214
- Pienemann, M and M Johnston 1986. 'An acquisition-based procedure for second language assessment'. *Australian Review of Applied Linguistics*, 9: 92–122
- Ross, S 1997. May. 'CSWE outcomes: 15 issues of description and inference'. Paper presented at *NCELTR National Forum on Assessment and Reporting in English Language and Literacy Programs: Trends and Issues*
- Ross, S 1998. 'Individual difference factors in CSWE outcomes: Implications for deeming criteria'. Unpublished manuscript
- Sasaki, M 1996. *Second language proficiency, foreign language aptitude, and intelligence*. New York: Peter Lang
- Schmidt, R 1983. Interaction, acculturation and the acquisition of communicative competence. In N Wolfson and E Judd (eds). *Sociolinguistics and second language acquisition*. Rowley, MA: Newbury House, 137–74
- Schumann, J 1978. *The pidginization process: A model for second language acquisition*. Rowley MA: Newbury House
- Scovel, T 1988. *A time to speak: A psycholinguistic inquiry into the critical period for human speech*. New York: Newbury House
- Singleton, D 1989. *Language acquisition: The age factor*. Clevedon: Multilingual Matters
- Skehan, P 1989. Individual differences in second language learning. London: Edward Arnold
- Skehan, P 1991. 'Individual differences in second language learning'. *Studies in Second Language Acquisition*, 13: 275–8
- Stern, H H 1976. 'Optimal age: myth or reality?' *The Canadian Modern Language Review/La Revue canadienne des langues vivantes*, 32: 283–94
- Walberg, H 1978. 'English acquisition as a diminishing function of experience rather than age'. *TESOL Quarterly*, 15: 427–37

Appendix 1

Australian Second Language Proficiency Ratings

Ingram and Wylie 1984

Abbreviations

S = speaking

L = listening

W = writing

R = reading

L1 refers to the native language

L2 or second language refers to the non-native target language

wpm = words per minute

Note: The authors acknowledge their debt to the FSI scale in the initial development stages of the ASLPR.

Zero proficiency

General Description

S:0 Zero proficiency

Unable to function in the spoken language.

Oral production is limited to, at most, occasional isolated words. Essentially no communicative ability

L:0 Zero proficiency

Unable to comprehend the spoken language.

Essentially no comprehension of even the most simplified and slowed speech.

W:0 Zero proficiency

Unable to function in the written language.

Essentially unable to communicate in writing, even though, if the L1 uses the same alphabet as the L2, may be able to form the letters and copy word shapes.

Comment

Learners at this level could include persons unable to read or write in any language, persons able to read or write in a language (other than the L2) which does not use the Roman alphabet, or persons able to read or write in a language (other than the L2) which uses a Roman alphabet.

R:0 Zero proficiency

Unable to comprehend the written language.

Essentially no comprehension of even isolated words or simple phrases.

Comment

Learners at this level could include persons unable to read or write in any language, persons able to read or write in a language (other than the L2) which does not use the Roman alphabet, or persons able to read or write in a language (other than the L2) which uses a Roman alphabet.

Initial proficiency

S:0+ Initial proficiency

General Description

Able to operate only in a very limited capacity within very predictable areas of need. Vocabulary limited to that necessary to express simple elementary needs and basic courtesy formulae. Syntax is fragmented, inflections and word endings frequently omitted, confused or distorted and the majority of utterances consist of isolated words or short formulae. Utterances rarely consist of more than two or three words and are marked by frequent long pauses and repetition of an interlocutor's words. Expression is often excessively marked by culturally inappropriate non-verbal features and sympathetic noises. Pronunciation is frequently unintelligible and is strongly influenced by first language. Can be understood only with difficulty even by persons such as officials or teachers who are used to speaking with non-native speakers or in interactions where the context strongly supports the utterance.

Examples of specific tasks (ESL)

Can give own name, age, address, phone number, number of children, nationality, ethnic group or country of origin of the family, and name of first language. Can use some basic greetings; can say *yes, no, pardon, excuse me, please, thank you, sorry*. Can spell out own name and those of family. Can make simple purchases where pointing or other gesture can support the verbal reference.

Comment

Areas of need may be those the learner experiences daily or that are regularly encountered through the objectives and teaching situations in the course followed. Interference from socio-cultural factors is particularly marked at this level and may inhibit language performance (eg if the L1 culture regards contradiction as impolite, yes-no questions may always be answered *yes* even though the correct answer is *no* and the *no* answer form has been mastered). Distortion of word endings may involve omission, addition or substitution of phonemes or allophones.

The ability to differentiate between surname, given names etc will depend on previous learning experiences and need not reflect upon language proficiency.

L:0+ Initial proficiency

General Description

Able to comprehend only a very restricted range of simple utterances within the most predictable areas of need and only in face-to-face situations with people used to dealing with non-native speakers.

Can comprehend only slow, deliberate speech in face-to-face situations. May require much repetition, paraphrase and the support of exaggerated mime and gesture. Commonly responds to isolated words in utterances. Misunderstandings are frequent and even short utterances must frequently be repeated. Able to comprehend responses to own simple questions pertinent to survival needs where those responses do not deviate far from the expected response. Can comprehend only very simple requests for basic information where context makes the nature of the request predictable and the forms closely match the formulae learnt. Can comprehend the simplest sentence structures only when supported by the context or other redundant features. Comprehends few non-verbal features not found in own culture. While listening, tends to make excessive use of 'sympathetic' features (eg nod, *yes*, *thank you*, repetition of speaker's words).

Examples of specific tasks (ESL)

Can comprehend memorised items in situations experienced in the learning environment. Can comprehend simple, predictable requests in predictable forms for personal and family information (name, date of birth, country of origin, language, telephone number, occupation). Less predictable questions are frequently interpreted as statements despite structure, intonation and context. Can comprehend basic time modifiers (eg *today*, *tomorrow*) and days of the week. Can comprehend basic directions such as *turn left* or *turn right*. Can comprehend two-word numbers (eg *thirty-six*) or three-digit numbers with numerals isolated (eg *one six four*), can comprehend high frequency sums of money provided that each unit (dollars, cents) does not exceed this complexity (eg ten dollars thirty-five). Can comprehend nod or shake of head, indication of direction, (pointing, beckoning), extension of hand.

W:0+ Initial proficiency

General Description

Able to write clearly a limited number of words or short formulae pertinent to the most predictable areas of everyday need.

Can provide basic information of immediate relevance in isolated words or short formulae. Can write with reasonable phonetic accuracy short words that are in his or her oral vocabulary. Has sufficient memory for word shapes to write recognisably (if not with formal accuracy) short words pertinent to written needs.

Examples of specific tasks (ESL)

Can copy names of everyday objects, names of shops and street signs. Can write own name, address, age, date of birth or arrival in the country, and those of family. Can write short, familiar, mainly one-syllable words when said aloud by self or others, not necessarily with correct spelling but with reasonable phonetic accuracy.

Comment

By *writing* is intended the ability to transfer sound into script. Whether acceptable cursive writing as well as print can be used at the lowest proficiency levels will vary according to the learner's background. Ability to read cursive writing will depend on similar factors but also on the writer's letter formation.

R:0+ Initial proficiency

General Description

Able to read only a limited range of essential sight words and short simple sentences whose forms have been memorised in response to immediate needs.

Where the language has an alphabet, can recognise most printed letters; can comprehend commonly encountered names and other isolated words whose forms have been memorised and which are relevant to everyday needs; can read most of them aloud, but comprehensibility may suffer because of sketchy knowledge of sound-symbol correspondence and inaccurate articulation. Is generally unable to recode unfamiliar words into sound except where considerable transfer from L1 to L2 is possible.

Examples of specific tasks (ESL)

Can identify the names of own family and place of living; can recognise names of common shops and familiar street signs (eg *Keep Left, Keep Right, Walk, Don't Walk*). Can identify and read aloud the names of common everyday objects as learnt in response to survival needs, though not necessarily with correct pronunciation.

Comment

Reading involves, most fundamentally, obtaining meaning from script, but related developments include sound-symbol correspondence and recoding word or sentence shapes into sound. At this level the beginnings of sound-symbol correspondence and visual memory for word shapes are emerging. The length of time needed to reach them will vary according to the learner's previous level of literacy, the scripts (if any) in which literacy exists, and the transferability of the sound-symbol relationships from the L1 to the L2.

Elementary proficiency

S:1- Elementary proficiency

General Description

Able to satisfy immediate needs using learned utterances.

The first signs of spontaneity and flexibility are emerging but there is no real autonomy of expression; there is a slight increase in utterance length but frequent long pauses and repetition of interlocutor's words still occur. Can ask questions or make statements with reasonable accuracy only where this involves short memorised utterances or formulae. Most utterances are telegraphic and word endings (both inflectional and non-inflectional) are often omitted, confused or distorted. Vocabulary is limited to areas of immediate survival needs. Can differentiate most phonemes when produced in isolation, but when they are combined in words or groups of words, errors are frequent and, even with repetition, may severely inhibit communication even with persons used to dealing with such learners. Little development in stress and intonation is evident. May make inappropriate use of sympathetic noises and gesture.

Examples of specific tasks (ESL)

Can produce utterances in fragmentary grammar which may consist of, eg no more than noun, verb and modifier. Can give in fragmentary utterances

and with much repetition such personal details as name, address, nationality, marital status, occupation, date and place of birth, date of arrival in country or other event. Can not use English over the telephone except in the most limited contexts where the information given or request made is highly predictable (eg to request an interpreter from the Telephone Interpreter Service). Can make simple purchases by stating what is wanted and asking the price, but only where the context supports the verbal message. Can buy tickets on public transport using utterances such as *Two returns Central, please*. Can indicate time by such phrases as *next week, last Friday, in November, three o'clock, two-twenty*. Can spell out name and address.

Comment

The extent of use of sympathetic noises and gesture is heavily influenced by cultural background and personality.

L:1- Elementary proficiency

General Description

Able to comprehend readily only utterances which are thoroughly familiar or are predictable within the areas of immediate survival needs.

In less familiar utterances, still tends to identify and respond to isolated words which will often cause misunderstanding and inappropriate responses. Still heavily dependent on face-to-face contact. Although able to recognise many basic patterns of structure and intonation, tentative grammatical knowledge will still cause many sentence meanings to be confused. Can comprehend only the commonest nonverbal features and has a developing awareness of some variants of them.

Examples of specific tasks (ESL)

Can comprehend requests for personal details (name, age, address etc) and short statements about others. Frequently interprets questions as statements unless repeated and redundantly marked by sentence structure, *Wh*-word, strong intonation, or context. Comprehends commands, requests and simple directions relating to how to get from X to Y only when supported by obvious nonverbal features. Cannot comprehend telephone, radio or television. Can comprehend three digit numbers if said deliberately (eg 2 3 6) and sums of money of equivalent difficulty (eg \$159.95). Comprehends basic time phrases such as *2.20, 3 o'clock, next week, in November*. Readily comprehends simple

past present and future tenses only if supported by an appropriate modifier. Can comprehend minimal variants of basic non-verbal features such as assent, negation or indication of direction.

Comment

Ability to comprehend sums of money is strongly influenced by cultural, vocational, personality and situational factors.

W:1- Elementary proficiency

General Description

Able to write with reasonable accuracy short words and brief familiar utterances.

Where the language has an alphabet, can form all letters; can write with reasonable phonetic and formal accuracy basic personal details and the names of everyday objects. Can write a short simple sentence or brief instruction relating to matters in areas of immediate need or with which he or she is very familiar.

Examples of specific tasks (ESL)

Can write short, familiar words with reasonable accuracy though may need to sound them out. Can write a phrase or short sentence (not necessarily accurately) to give basic details about self and family, to reply to a query, or to convey simple information (eg to identify a photograph or a familiar scene). Can copy short written sentences (eg giving directions how to go from X to Y).

R:1- Elementary proficiency

General Description

Able to read short simple sentences and short instructions.

Can recognise and name most of the letters of the printed alphabet (both upper and lower case if found in the language). Can read short, original sentences of one clause on familiar topics. Fluency is restricted by syntactic knowledge, vocabulary, cultural knowledge and inability to handle longer sentences or the discourse structure of even short texts. Silent reading may be accompanied by oral recoding. Word recognition may depend heavily on sounding out the letters. Aware of the more frequent sound-symbol correspondences, but errors will still occur frequently.

Examples of specific tasks (ESL)

Can identify and read aloud common words (eg names of shops and everyday objects). Can read aloud with reasonable accuracy words containing familiar sound-symbol correspondences, though stress-patterns may be faulty. Can comprehend short written directions (eg to go from X to Y) or simple one-sentence instructions (eg arranging time and place to meet). Can understand commonest abbreviations in daily usage such as *Mr*, *Mrs*, *am* and *pm*.

Minimum survival proficiency

S:1 Minimum survival proficiency

General Description

Able to satisfy survival needs and minimum courtesy requirements.

In areas of immediate need or on very familiar topics, can ask and answer simple questions, initiate and respond to simple statements, and maintain very simple face-to-face conversations. Vocabulary inadequate to express anything but the most elementary needs; fractured sentence structure and other grammatical errors are frequent; strong interference from L1 occurs in articulation, stress and intonation. Misunderstandings frequently arise from limited vocabulary and grammar and erroneous phonology, but, with repetition, can generally be understood by native speakers in regular contact with foreigners attempting to speak their language. Little precision in information conveyed owing to tentative state of grammatical development and little or no use of modifiers. Has not developed skills in a specialist register; though, where such a register has been experienced, may have acquired some relevant items.

Examples of specific tasks (ESL)

Despite hesitations, fractured syntax and many repetitions, can give personal information and maintain very simple conversations on topics that are familiar or of personal interest; can express likes and dislikes in areas of particular interest; can make an introduction and use basic greeting and leave-taking expressions; can ask and tell time of day, day and date- can make simple transactions in shops, post offices or banks; can verbalise inability to understand, ask for slower repetition of utterance, spelling of name or address. Depending on need and previous experience, can order a simple meal, ask for shelter or lodging, ask and give simple directions (eg tell someone how to get from X

to Y); can use public transport (buses, trains, and taxis, ask for basic information, ask and give directions, and buy tickets).

Comment

Modifying devices or modifiers are those forms (eg verb forms, adjectives, adverbs, phrases, clauses etc) that are used to modify and qualify ideas and give precision to the expression of thought.

L:1 Minimum survival proficiency

General Description

Able to comprehend enough to meet basic survival needs.

Can comprehend short utterances and some longer ones provided the content is familiar to him or they are in response to his own simple questions pertinent to personal details or survival needs in his everyday life. Only common social formulae or other short, simple, familiar utterances are readily understood; otherwise, requires frequent repetition, redundancy, or paraphrase and a slow deliberate rate of utterance. Lacks discourse mastery and generally unable to relate even a short series of utterances. Has very limited ability to cope with specialist registers though may comprehend some items pertinent to areas of activity which have been experienced.

Examples of specific tasks (ESL)

Can comprehend simple directions relating to how to get from X to Y, by foot or public transport; can comprehend time of day, day and date, and appointments; can cope with only simple verbal number operations. Can respond to simple high frequency instructions in familiar situations (eg school or work); comprehends less predictable utterances in familiar situations if said slowly and deliberately (eg 80–100 wpm). Has great difficulty in using a telephone or comprehending radio or television. Comprehends only the most frequently occurring contracted forms (eg *I'm, it's, don't, can't, won't, isn't*). Is, to only a limited extent, sensitive enough to suprasegmental, non-verbal and other paralinguistic features to discriminate 'tone' of utterances (eg polite, rude, friendly, unfriendly etc.).

Comment

Familiar situations could include those regularly encountered in work, school, leisure etc.

W:1 Minimum survival proficiency

General Description

Able to satisfy basic survival needs.

Can write all letters of the alphabet and copy most sentences accurately. Can write personal details and a short series of sentences about things that are familiar (not necessarily with formal accuracy in lexis and syntax but comprehensibly). Longer utterances or longer series of utterances tend to lose coherence.

Examples of specific tasks (ESL)

Can fill in uncomplicated forms with personal details (name, address, nationality, marital status). Can write simple sentences, including brief instructions (eg to door-to-door vendors). Can copy from written script quite accurately the sorts of information indicated in R:1. Can copy down, when presented orally, name, address, appointment and simple directions. Can fill out bank deposit and withdrawal forms. Can produce a short series of simple sentences on a familiar topic (eg personal details for a job application, short simple narration of an everyday occurrence, postcard).

R:1 Minimum survival proficiency

General Description

Able to read personal and place names, street signs, office and shop designations, numbers, isolated words and phrases, and short sentences.

Can recognise and name all the letters in the printed version of the alphabet (both upper and lower case if found in the alphabet). Can read simple sentences with ease, but may have difficulty with sentences of greater complexity. Can read and comprehend (but not necessarily fluently) very short, syntactically simple texts concerning his everyday life (cf L:1), but has considerable difficulty in comprehending texts with more complex discourse structure.

Examples of specific tasks (ESL)

Can read uncomplicated forms requiring basic personal details (name, address, nationality, marital status) but will require help in comprehending others. Can comprehend short, high frequency traffic signs, shop designations, bus and traffic destinations, basic timetables, and common English language menus. From lists (eg street directory, index) can isolate the information required. Can

use the 'Yellow Pages' to find a tradesman. Can read a notice for a function and identify the nature, name, location, date and time of the event. Can read a short series of simple sentences narrating an everyday event. Can comprehend and act on a short series of simple instructions for using an everyday object (provided they do not contain unfamiliar specialist vocabulary).

Survival proficiency

S:1+ Survival proficiency

General Description

Able to satisfy all survival needs and limited social needs.

Developing flexibility in a range of circumstances beyond immediate survival needs. Shows some spontaneity in language production but fluency is very uneven. Can initiate and sustain a general conversation but has little understanding of the social conventions of conversation; grammatical errors still frequently cause misunderstanding. Limited vocabulary range necessitates much hesitation and circumlocution. The commoner tense forms occur but errors are frequent in formation and selection. Can use most question forms. While basic word order is established, errors still occur in more complex patterns. Can not sustain coherent structures in longer utterances or unfamiliar situations. Ability to describe and give precise information is limited by still tentative emergence of modification devices. Aware of basic cohesive features (eg pronouns, verb inflections), but many are unreliable, especially if less immediate in reference. Simple discourse markers are used relating to closely contiguous parts of the text, but extended discourse is largely a series of discrete utterances. Articulation is reasonably comprehensible to native speakers, can combine most phonemes with reasonable comprehensibility, but still has difficulty in producing certain sounds, in certain positions, or in certain combinations, and speech may be laboured. Stress and intonation patterns are not native-like and may interfere with communication. Still has to repeat utterances frequently to be understood by the general public. Has very limited register flexibility, though, where a specialist register has been experienced, may have acquired some features of it.

Examples of specific tasks (ESL)

Can cope with less routine situations in shops, post office, bank (eg asking for a larger size, returning an unsatisfactory purchase), and on public transport (eg

asking passenger where to get off for unfamiliar destination). Can explain some personal symptoms to a doctor but with limited precision. Can modify utterances to express uncertainty or the hypothetical by single word or other simple devices (eg *possibly, I think*) and has tentative use of if (conditional). Can use simple discourse markers such as *so, but, then, because*. Often makes inappropriate use of honorifics eg title without surname. In work situation can communicate most routine needs not requiring special register (eg out of expendable commodity or a machine overheating) and basic details of unpredictable occurrences, eg an accident. Can ask the meaning of an unfamiliar word, or ask for the English word for a demonstrable item. Can generally use *I, me, you, we, my, your*, but other personal pronouns and possessive adjectives are often hesitant or wrong.

Comment

From this level on, the learner has a significant language repertoire permitting comprehension of texts containing an increasing number of unfamiliar language items or cultural references. The learner now has a sufficient language base to benefit greatly in language learning from out-of-class experience and to permit exploration of the language by inquiry from native speakers. The thrust of development through this level is towards more spontaneity and creativity, increased flexibility but still in essentially survival-type situations with a start to more general social interaction. Ability to comprehend still depends greatly on the native speaker's modifying the language produced. Immediate memory is less restricted, operations less laboured and some textual facility is starting to emerge. Cultural interference may create unease in use of second person pronouns and persons' names for learners of some backgrounds. Some pronunciation problems will persist well beyond this stage. The ability to acquire flexibility in social register varies greatly according to the background, sensitivity and personality of the individual. The emergence of modifying devices and discourse markers gives the learner the means to express (however tentatively at 1+) individual meanings (eg personal perceptions and attitudes) as well as universal meanings. 'Work' situations should be considered to include school for students.

L:1+ Survival proficiency

General Description

Able to satisfy all survival needs and limited social needs.

Can understand in all situations relevant to his or her survival needs. Less dependent on contextual support but comprehension is still significantly assisted by face-to-face contact, careful articulation and slow rate of utterance. Provided the topic discussed is familiar, can extrapolate the meaning of occasional unknown words from the context and deduce sentence meaning. While global comprehension of utterances may be generally secure, deficiencies in the listening skill will often lead the learner to miss more specific information; misinterpretations are frequent and, in less familiar situations, may require repetition, paraphrase or explanation. More sensitive to morphology, recognises changes of tense and generally comprehends regular forms without having to rely on contextual support. Has much difficulty comprehending extended lines of argument and comprehends only the simplest discourse markers. Has little ability in specialist registers but can follow routine communications in the workplace and some unpredictable utterances provided they involve mainly the non-specialist register or are supported by the context. Can generally distinguish statements, questions and commands by intonation etc, but has limited ability to deduce other than surface meaning. Comprehension still suffers from limited familiarity with the target culture.

Examples of specific tasks (ESL)

Recognises simple relationships between short combinations of clauses and sentences (eg marked by and, but, if, because), but fails to comprehend more complex relationships or more complex discourse marked over a longer text. Can cope with utterances that are carefully articulated and said slowly (eg 100 to 120 wpm.). Comprehends only isolated words or phrases in most conversations between native speakers and will also fail to comprehend subsumed knowledge. Has no facility in comprehending speakers of dialects other than that most frequently experienced. Comprehends most common, standard contracted forms but has considerable difficulty with colloquial 'run-on' forms (such as *wanna, gunna, wotcha*). Comprehends very little of a radio broadcast and has great difficulty using the telephone. Can comprehend only broad train of events of a TV drama little of less visually supported telecasts, though has fair comprehension of frequently repeated commercials said slowly in the standard dialect without significant cultural assumptions. If heavily marked,

understands the implied annoyance, sarcasm etc in such utterances as *Haven't you finished yet?* or *That's great!*. Can act on a simple series of commands such as *Find a rag and clean the bench* and negative commands such as *Don't touch that button*.

Comment

From this level on, the learner has a significant language repertoire permitting comprehension of texts containing an increasing number of unfamiliar language items or cultural references. The learner now has a sufficient language base to benefit greatly in language learning from out-of-class experience and to permit exploration of the language by inquiry from native speakers. The thrust of development through this level is towards more spontaneity and creativity, increased flexibility but still in essentially survival-type situations with a start to more general social interaction. Ability to comprehend still depends greatly on the native speaker's modifying the language produced. Immediate memory is less restricted, operations less laboured and some textual facility is starting to emerge. Cultural interference may create unease in use of second person pronouns and persons' names for learners of some backgrounds. Some pronunciation problems will persist well beyond this stage. The ability to acquire flexibility in social register varies greatly according to the background, sensitivity and personality of the individual. The emergence of modifying devices and discourse markers gives the learner the means to express (however tentatively at 1+) individual meanings (eg personal perceptions and attitudes) as well as universal meanings. 'Work' situations should be considered to include school for students.

W:1+ Survival proficiency

Able to satisfy all survival needs and limited social needs.

Can write a sufficiently wide range of informal language to satisfy survival needs though errors in syntax, spelling and style may frequently interfere with comprehension. Can write most words and sentences that can be produced orally but longer utterances and texts may lack coherence through a failure to maintain sentence structure, to structure the thought sequence acceptably, or to use discourse markers appropriately. Has the ability to use a bilingual dictionary to check spelling. Aware of but not confident with the formal devices used in writing letters and has limited ability to vary them to match different recipients.

Examples of specific tasks (ESL)

Can write a simple covering letter (eg to accompany a cheque or a completed job application form). Can convey a simple message by telegram (eg accepting job offered) though the conventions and form may often be inappropriate. Can write a note to school explaining a child's absence or requesting leave for the child. Can take down a simple message in note form.

R:1+ Survival proficiency

Able to read short texts on subjects related to immediate needs.

Ready comprehension is conditional on the meaning being clearly spelt out. Can follow a simple compound or complex sentence, but has little ability to handle many discourse markers commonly used in the written registers. With some use of a bilingual dictionary can read for pleasure simplified versions of standard texts. Word attack is now sufficiently developed for the learner to be able to recognise most words in oral vocabulary though silent letters, non-phonetic forms or other irregularities may still cause confusion. Where texts are in cursive writing, copes only if the writing is neat and the style familiar.

Examples of specific tasks (ESL)

Can understand a simple circular sent home from school on a familiar topic (eg start of swimming season) or an unfamiliar topic (eg an excursion), provided that some prior explanation has been given. Using a bilingual dictionary can read a popular novel or short story simplified for L2 learners. Can comprehend texts in which discourse is marked by such simple features as those indicated in S:1+. Can follow the instructions on a public telephone provided such words as *dial tone* and *receiver* are familiar. Can discriminate between widely differing letters such as an apparently personalised sales promotion and correspondence requiring action (eg overdue account or alteration to delivery routine) even if neither is fully understood. Can understand straightforward classified advertisements in which the information is directly stated, which do not use unfamiliar registers, contain a high degree of abbreviation, or depend significantly on the learner's ability to supply implied meaning (eg as in innuendo).

Minimum social proficiency

S:2 Minimum social proficiency

General Description

Able to satisfy routine social demands and limited work requirements.

Can handle with confidence but not facility most social situations including introductions and casual conversations about current events, as well as work family and autobiographical information. Has restricted register flexibility though, where a specialist register has been experienced, will have acquired some features of it. Has limited ability to vary the 'tone' of utterances. Can Handle limited work requirements but will need help in handling any complications or difficulties. Hesitations are still frequent as the learner searches for vocabulary or grammar, but has a speaking vocabulary sufficient to express himself simply with circumlocutions on most topics pertinent to his everyday life; can usually handle elementary constructions quite accurately but does not have thorough or confident control of the grammar especially in longer constructions. Accent, though often quite faulty, is intelligible, undue exertion on the part of a native-speaking listener is not often necessary though some repetition in order to be understood may occur. Overall rate of utterance remains less than the native speaker's as a result of hesitations Cohesion and discourse in short utterances or texts are secure but inconsistencies occur in longer ones.

Examples of specific tasks (ESL)

Can give detailed information about own family, living conditions, educational background; can describe and converse on everyday things in his environment (eg his suburb, the weather); can describe present or most recent job or activity; can communicate on the spot with fellow workers or immediate superior (eg ask questions about job, make complaints about work conditions, time off etc.); can give simple messages over the telephone; can give directions and instructions for simple tasks in his everyday life (eg to tradesmen). Has tentative use of polite request forms, eg involving *could*, *would*. May sometimes offend by unintended blandness or aggressiveness, or irritate by over-deference where native speakers expect informality.

Comment

At this level, the learner's ability is sufficient to enable him to establish normal social relationships with native speakers.

L:2 Minimum social proficiency

General Description

Able to understand in routine social situations and limited work situations.

Can get the gist of most conversations in everyday social situations though may sometimes misinterpret or need utterances to be repeated or reworded. Less dependent on face-to-face contact, and the presence of other participants in the conversation does not normally cause comprehension problems. Has limited ability to comprehend if there is extensive use of specialist registers though, in own field, can, with paraphrase or explanation, comprehend routine conversations. Has some ability to see beyond surface meaning to comprehend less subtle or esoteric cultural implications. Can cope with the lower range of normal native speaker utterance rates, but is soon lost with faster rates and has difficulty in understanding conversations between native speakers. Has some difficulty following extended lines of argument or complex discourse patterns.

Examples of specific tasks (ESL)

Can readily understand the sort of information indicated in S:2 when given face-to-face; can take simple telephone messages in response to own questions or on familiar or expected topics; can cope with native speaker conversations at lower rates of utterance (eg 120–150 wpm). Can comprehend and act on sequential instructions at work (eg *when the red light goes off, push the button*); does not fully comprehend television or radio broadcasts, but can get the gist of news bulletins or other programs on familiar topics delivered at lower rates of utterance. Can, in most situations, broadly discriminate the ‘tone’ of utterances (eg irony).

W:2 Minimum social proficiency

General Description

Able to satisfy routine social demands and limited work requirements.

Can handle (with moderate confidence and sufficient accuracy that comprehension by a native speaker is not impeded) those written forms needed in his or her everyday life at home, in daily commerce and in simple work situations not requiring specialist skills.

Examples of specific tasks (ESL)

Can write a personal letter on simple everyday topics or a simple report on an everyday event. Can write to order goods, book a room, or to carry out other uncomplicated and routine tasks. Can fill out most forms regularly encountered in everyday life (eg health insurance, unemployment registration, passport application etc).

R:2 Minimum social proficiency

General Description

Able to read simple prose, in a form equivalent to typescript or printing, on subjects within a familiar context.

With extensive use of a bilingual dictionary can get the sense of those written forms frequently met in his or her everyday life. Can read for pleasure simply structured prose and literary and other texts which do not assume significant cultural knowledge, ability to handle complex discourse structure, or a specialist register. Can read neat cursive writing if the style is familiar.

Examples of specific tasks (ESL)

Using a bilingual dictionary, can get the sense of personal letters on everyday topics, simple stylised forms such as invitations and replies, routine, uncomplicated business letters, news items from the daily press (but longer reports and commentaries only with considerable difficulty), and simple articles in technical fields relevant to work experience. Can follow most clearly presented sequential instructions (eg accompanying a household appliance) when they are written in a non-specialist register, when there is plenty of time and a bilingual dictionary is available. Can read fluently for pleasure modern novels simplified for the non-native reader.

Minimum vocational proficiency**S:3 Minimum vocational proficiency**

General Description

Able to speak the language with sufficient structural accuracy and vocabulary to participate effectively in most formal and informal conversations on practical social and vocational topics.

Can discuss own particular interests and special fields of competence with rea-

sonable ease though some circumlocutions, vocabulary is broad enough that the learner rarely has to grope for a word and can readily overcome gaps with circumlocutions; accent may be obviously foreign; control of grammar good; able to convey meaning precisely in reasonably complex sentences or by using, with reasonable accuracy, a wide range of modification devices; fluency is rarely disrupted by hesitations; errors rarely interfere with understanding or disturb the native speaker; able to modify language to meet the differing register requirements of situations which are familiar in the learner's personal and vocational life but can make secure use of only high frequency colloquialisms.

Examples of specific tasks (ESL)

Can handle with confidence most social situations and those work situations relevant to own needs and experience. Can enter, exit from and participate in conversation with or between native speakers; can speak to educated native speakers or to those at own socioeconomic level on general or relevant vocational topics without unintentionally amusing or irritating them; can present and debate own or others' ideas and attitudes about familiar topics or topics which are relevant to own or target culture, can cope with everyday difficult linguistic situations, such as broken plumbing, a personal misunderstanding, undeserved traffic ticket etc.

Comment

The key factor now emerging is register flexibility (as well as continued development in fluency and accuracy) Fluency refers to the ability to mobilise language components in connected expression.

L:3 Minimum vocational proficiency

General Description

Able to comprehend sufficiently readily to be able to participate effectively in most formal and informal conversations with native speakers on social topics and on those vocational topics relevant to own interests and experience.

Can get the gist of most conversations between native speakers though may miss some details, especially where there is significant subsumed knowledge. Comprehension rarely affected by complex discourse patterns. Can generally understand at normal rates of utterance, even if occasional words are unfamiliar, and rarely has to ask for an utterance to be repeated or paraphrased, except where speech is heavily loaded with colloquial features.

Examples of specific tasks (ESL)

Can take information confidently by telephone; can comprehend a discourse or discussion on a non-technical subject and can, if necessary, take notes or summarise it, has reasonable comprehension of radio and television news readers (180 wpm) though more rapid speakers may cause comprehension to suffer. Can comprehend most sums of money and most numerals though longer items may have to be repeated.

Comment

It should be noted that colloquial speech may entail changes in every aspect of language, including vocabulary, syntax, semantics, phonology, rate of utterance and paralinguistics.

W:3 Minimum vocational proficiency

General Description

Able to write with sufficient accuracy in structures and spelling to meet all social needs and basic work needs.

Can write with reasonable ease and accuracy on matters relevant to own interests, rarely lacks a word and is then able to circumvent it. Can use complex sentences accurately and can vary the style between personal and vocational contexts and use the functions appropriate to them.

Examples of specific tasks (ESL)

Can write in all those forms used in daily life (personal letters, notes, telegrams, invitations etc) without errors intruding on a native speaker's comprehension and acceptance. Can use those basic registers needed in the work situation and other routine business letters of his or her everyday life. Can generally use even complex sentence structures accurately. Discourse structure beyond the sentence level may still sometimes seem non-native. Can vary style over broad parameters (eg between personal and vocational contexts), though may lack some subtlety in differentiating between some contexts.

Comment

Principal changes from here to R-5, W-5 lie in breadth of vocabulary, accuracy of syntax and flexibility of, or sensitivity to, style.

R:3 Minimum vocational proficiency

General Description

Able to read standard newspaper items addressed to the general reader, routine correspondence, reports and technical material in own special field, and other everyday materials (eg best-selling novels and similar recreational literature).

Can grasp the essentials of articles of the above types without using a dictionary; for accurate understanding, moderately frequent use of a dictionary is required. Has occasional difficulty with unusually complex structures and low-frequency idioms. Can read cursive writing though non-standard or ill-formed scripts may still cause difficulty.

Examples of specific tasks (ESL)

Can read standard newspaper items and routine personal and business correspondence in own field of interest and readily grasp their essential meaning. Can read such items and paraphrase or summarise their key points. Has some sensitivity to variations in style and register and to nuances in meaning but will frequently fail to perceive those that are more subtle or more culturally dependent. Is able to read extended texts (eg novels) with sufficient comprehension without reference to a dictionary to ensure pleasure even though some words will be unknown.

Vocational proficiency**S:4 Vocational proficiency**

General Description

Able to use the language fluently and accurately on all levels normally pertinent to personal social academic or vocational needs.

Can participate in any conversation within the range of own experience with a high degree of fluency and precision of vocabulary; while the learner has mastered commonly occurring colloquial and idiomatic forms, some misuse of other items may occur; would rarely be taken for a native speaker, but can respond appropriately even in unfamiliar situations; while a 'foreign accent' may continue (especially in intonation and stress patterns), pronunciation does not impede comprehension by a native speaker; errors of grammar are quite rare and unsystematic and can usually be corrected in retrospect; always easily

understood by a native speaker. Has considerable sensitivity to register requirements and readily modifies the language appropriately.

Examples of specific tasks (ESL)

Can convey exact meaning in social and vocational discussions unrestricted by lexical or grammatical deficiencies; can modify speech deliberately according to the situation and its register requirements can handle informal interpreting from first language.

Comment

Cultural understanding now plays a significant part in promoting language use. Grammatical development is now more or less complete though 'slips' or errors of performance may still occur. The learner can, however, usually correct such errors if he or she becomes conscious of them.

L:4 Vocational proficiency

General Description

Can comprehend easily and accurately in all personal and social contexts and in all academic or vocational contexts relevant to own experience.

Can readily understand all speech of that variety of the language normally encountered in own personal, social, academic or vocational life and is rarely troubled by speech in less familiar contexts; can comprehend even fast rates of utterance in the target variety- can comprehend the generally recognised varieties of the target language and other similar varieties; only occasionally baffled by colloquialisms and regionalisms.

Examples of specific tasks (ESL)

Can comprehend accurately in social and vocational discussions, unrestricted by lexical or grammatical deficiencies; can comprehend even fast utterance rates of 180 to 200 or more wpm in the target variety; can appreciate and respond to register variations. Can comprehend most radio and television documentaries, and accurately identify the speaker's mood, tone etc., can comprehend numerals as readily as a native speaker. If learning Australian English, can comprehend such varieties as Educated, General and Broad Australian, RP, Educated Indian and N.E. American, but will have increasing difficulty with more distant varieties. Can handle informal interpreting into first language.

W:4 Vocational proficiency

General Description

Able to write fluently and accurately on all levels normally pertinent to personal, social, academic or vocational needs.

Errors in grammar or vocabulary are very rare, rarely needs to consult a dictionary to express himself and is able to consider and select from amongst a wide choice of words and structures to make meaning more precise. Has considerable sensitivity to register requirements and can modify language appropriately.

Examples of specific tasks (ESL)

Can convey meanings precisely and accurately unrestricted by lexical, morphological, syntactic or spelling deficiencies. Can readily use all those written forms normally encountered and can modify them according to specific register requirements. Can structure longer texts appropriately, making full use of the available devices of discourse and cohesion. Can handle informal translation from first language.

R:4 Vocational proficiency

General Description

Able to read all styles and forms of the language pertinent to personal, social, academic or vocational needs.

With occasional use of a dictionary can read moderately difficult prose readily in any area directed to the general reader, and all material in own special field including official and professional documents and correspondence. Reading speed will approximate that of comparably educated native speakers. Cursive writing poses no greater difficulty than for a native speaker.

Examples of specific tasks (ESL)

Can comprehend most literary forms though more difficult works (eg heavily culture-dependent or in a form remote from 'normal' discourse) may cause some problems. Can handle informal translation into first language.

Native-like proficiency**S:5 Native-like proficiency**

General Description

Speaking proficiency equivalent to that of a native speaker of the same socio-cultural variety.

Has complete fluency, accuracy and range in the language such that the learner's speech on all levels is fully accepted by such native speakers in all its features (including rate of utterance, suprasegmental and paralinguistic features, and breadth and accuracy of grammar, vocabulary, idiom, colloquialisms and cultural references), even though some phonological features may exhibit minor non-native characteristics that never intrude nor inhibit comprehensibility. Able to operate as effectively as a native speaker of the same socio-cultural variety in all those registers encountered in his personal, social, academic or vocational life.

Examples of specific tasks (ESL)

Can handle all tasks normally encountered and has native-like flexibility in new ones; can handle humour and innuendo as effectively as a native speaker of the same socio-cultural variety. Very occasional non-native syllable stress, intonation pattern, allophone or phoneme substitution.

Comment

At this point all language-related cultural barriers are removed and to all intents and purposes, the learner acts, and is accepted by others, as a native speaker. The distinction between the terms variety and register should be recalled. Any register limitations are of the same linguistic order and cause as for a native speaker of the same socio-cultural variety. It should be remembered that the ASLPR measures general proficiency. Register flexibility is as firmly established at this level as for a native speaker. Learners who have not experienced a certain specialist register in their L2 may not have facility in it even though that register (eg of an academic or sporting interest) may be fully developed in their L1. With exposure to the register, however they will master it as readily as would comparable native speakers or, perhaps, more readily if they already have the relevant underlying concepts.

L:5 Native-like proficiency

General Description

Listening proficiency equivalent to that of a native speaker of the same socio-cultural variety.

Has the same degree of comprehension of the spoken language of native or non-native speakers in all its features (including idiom, colloquialisms, subtlety of meaning, and cultural references whether spoken face-to-face, by telephone or on the media) as a native speaker of his socio-cultural variety. Can comprehend fully all varieties and registers likely to be encountered in his personal social, academic or vocational life and have sufficient flexibility to comprehend others as effectively as do native speakers of his socio-cultural variety. Similarly sensitive to the implications of the variety and register used.

Examples of specific tasks (ESL)

Can perform as effectively as a native speaker in all listening activities encountered and has a native speaker's flexibility in new ones. Perceives and responds to style, humour and innuendo as effectively as a native speaker of the same socio-cultural variety. Can generally recognise the likely educational level and origin of a speaker.

W:5 Native-like proficiency

General Description

Written proficiency equivalent to that of a native speaker of the same socio-cultural variety.

The learner's written language in all its forms is fully accepted by such native speakers in all its features including formal accuracy, structural variation, word choice, idiom, colloquialisms, register appropriateness, discourse structure (including thought sequence and coherence), subtlety of meaning and cultural references. Deviations from educated native speaker forms, special register features or stylistic conventions will only be those recognisable as native speaker variants.

Examples of specific tasks (ESL)

Can perform as effectively as a native speaker in all writing tasks normally encountered and has native-like flexibility in mastering new ones.

R:5 Native-like proficiency

General Description

Reading proficiency equivalent to that of a native speaker of the same socio-cultural variety.

Can comprehend all forms of the written language (including abstract, structurally complex, or highly colloquial literary and non-literary writings) as effectively as such a native speaker. Can comprehend fully all varieties and registers likely to be encountered in his or her personal, social, academic or vocational life and has sufficient flexibility to comprehend others as effectively as native speakers of the same socio-cultural variety. Is similarly sensitive to the implications of the variety and register used. Can appreciate humour and subtle or culture-dependent nuances of meaning or style. Has no more difficulty than a native speaker in reading handwriting and alternative scripts.

Examples of specific tasks (ESL)

Can comprehend with as much ease as an educated native speaker all forms of the written language normally encountered. Can identify the likely educational level of the writer. Can understand common references from the Judaeo-Christian literary tradition, and sport. Comprehends a text in Old English script as readily as a native speaker.

© Commonwealth of Australia

Appendix 2

ASPLR Reading Tasks

1. What is this story about?
2. What was Kim's problem when she first arrived in Australia?
3. Did she study English before? How much?
4. Did she study English before? How much?
5. What does she want to be able to do with her English?

My life in Australia

When I first arrived in Australia, I was afraid of everything. I couldn't speak any English or even understand what people said. Also, the way people lived was so different to the way we lived in Korea.

Now, six months later things are getting better, although I still worry about a few things.

The English language is my biggest problem. I studied English at high school in Korea for six years, for four hours a week, but my English still isn't good enough to join in conversations with Australians, to read books and magazines without using a dictionary, or to get the sort of job I want.

Kim Lee

1. Why did the CES send this letter to Mary?
2. Is 1pm a good time for her to call in?
3. What must Mary do if she cannot attend the interview?
4. When is the last possible date for an interview?
5. What will happen if she does not attend the interview or contact the CES?



Broadway Job Centre
818 George Street
Melbourne, Vic 3020

Phone 311-5020

OUR REF: 123-567112

REQUEST TO ATTEND A FULL INTERVIEW

Dear *Mary*.....

It is some time since you first registered with this office. We would now like to conduct a more detailed interview with you and discuss your job search and training options.

You are welcome to come into the office anytime from 8.45am to 5pm weekdays: although it is recommended that you do not call in between the hours of 12 noon and 2pm as you may experience some delays in service.

You must however come into this office for an interview before Wednesday the 28th August 1993.

If due to exceptional circumstances you are unable to call into this office before that date, please phone Debbie Fisher or Michael Gleave on 311-5020 so that other arrangements can be made.

If you don't come to the interview or call to make other arrangements, we will assume you no longer need CES help and cancel your registration.

PLEASE BE AWARE THAT IF YOU ARE ON JOB SEARCH ALLOWANCE OR NEWSTART ALLOWANCE, YOUR PAYMENTS WILL BE IMMEDIATELY CANCELLED AND A PENALTY WILL BE IMPOSED SHOULD YOU WISH TO RE-REGISTER FOR THE ABOVE ALLOWANCE.

[Signature]

Manager
Broadway Job Centre

13th August 1993

The Client Services Network of the
Department of Employment, Education and Training

A three-minute disaster

1. What is this story about?
2. How did the fire start?
3. What did the man do then?
4. What happened to the house?
5. How was the man injured?

A three-minute disaster

SYDNEY — A man burned his house down yesterday while trying to boil eggs.

The Wahroonga man was happily watching the prospective meal when he spotted a couple of cockroaches.

So he grabbed an insecticide pressure can and sprayed the offending beasts.

But some of the insecticide blew on to the gas flame and set alight the can and stove.

He then rushed into a bedroom, grabbed a blanket, ran back into the kitchen and tried to smother the fire.

When the blanket caught fire, the man tried to throw it out of the window, but it was closed.

He panicked and jumped through the window, glass and all.

As the flames from the burning insecticide can, stove, and blanket spread through the kitchen, he ran next door and phoned the fire brigade.

But by the time they got there, his \$120,000 northern suburbs home was ablaze and beyond saving.

The hapless cook was treated for cuts and shock.

Appendix 3

CSWE Reading Assessment Tasks

All tasks in Appendix 3 from *Certificates in Spoken and Written English III* (1995) with permission NSW Australian Migrant English Service (AMES).

Certificate in Spoken and Written English III (Vocational English) Assessment Competency 9: Can read an information text

Read the text then answer the questions following.

Job and course explorer

Job Profile — Clerk

Clerks are employed both in private industry and by the NSW and Australian Public Services. Depending on where they work, clerks may carry out all kinds of different duties. Some may work with numbers while others may write letters, or be involved with customer enquiries.

Duties

In many cases the work is in one particular area. The duties are specialised and clerks may be called by different titles depending on the duties they are required to perform. For example, an insurance clerk could deal with insurance claims — handling enquiries and issuing renewals; a personnel clerk could look after personnel details — keeping employees records up to date. Other clerical positions include accounts clerk or administrative services officer.

Clerical duties may include the following: filing and keeping files up to date; sorting and dispatching mail; maintaining recording systems and financial records; answering enquiries; writing letters and memos and a wide range of other duties. Due to the increasing mechanisation of office work, clerks may have to type, wordprocess documents, input data or code information for computers.

Personal requirements

Clerks require a methodical approach, attention to detail, an ability to communicate with the public for some jobs and keyboard skills for some positions.

Education and training

Employer requirements vary. Some may require completion of year 10 while others may prefer the HSC. Training is usually on-the-job; for some positions keyboard skills or previous clerical experience may be required

Employment opportunities

Clerks are employed in all industry sectors. The continuing introduction of new technology means that the clerical area is undergoing major changes and there is a continuing need to re-assess skill requirements. There is a ready supply of people with general clerical skills seeking work in this area.

Certificate in Spoken and Written English III (Vocational English) Assessment Competency 9: Can read an information text

1. What work does a personnel clerk do?

2. Machines are now more widely used in the workplace. What additional tasks may clerks now do as a result?

3. Give two important personal skills a clerk should have?
 - a. _____
 - b. _____
4. Where is most clerical training carried out?

5. Why are clerical skill requirements changing?

6. Why may it be useful, when looking for a clerical job, to have specialist clerical skills?

**Certificate in Spoken and Written English III (Vocational English)
Assessment Competency 9: Can read an information text**

Answer key: Job profile

1. look after personnel details — keeping employees' records up-to-date
(Need all above information)
2. type/wordprocess documents/input data/code information for computers
(Should have at least two of these answers)
3. methodical approach/and attention to detail/ability to communicate
(Any two)
4. on-the-job
5. introduction of new technology
6. there is ready supply of people with general clerical skills/better employment opportunities
(or words to that effect)

**Certificate in Spoken and Written English III (Vocational English)
Assessment Competency 9: Can read an information text**

Read the text then answer the questions following.

How to handle on-the-job PRESSURE

Most people have felt pressure, or stress, at some time during their working lives. On-the-job pressure comes in two forms. It can be pressure that arises from time to time because of a crisis. Some people don't handle this type of pressure well, while others find it exhilarating.

Or it can be the type of pressure which relates to constant urgency to complete more work than you can comfortably handle — always thinking that soon it will get easier, things will calm down and you'll have less on your plate, but never reaching that feeling of easy control. It is this type of pressure that is putting more and more people's health at risk.

Even though some people may enjoy the pace, the long days and accumulated fatigue can lead to physical and mental health problems.

Overwork creates health and safety problems because it eats into precious leisure time and leaves people lethargic, angry and unfocused. Many people work between 50 and 60 hours each week, which means they are living to work rather than working to live.

Some people find they are working longer hours, but not getting more done. When people are exhausted, their productivity declines. Tiredness can also adversely affect people's decision-making skills.

Coping with overload

Think about your priorities. How much time do you actually spend on doing the things you love? What changes can you make that will allow you to have more time available to do these things? Here are some suggestions.

- **Consider telling your supervisor that you aren't coping**
Your supervisor may be completely unaware of this and may be more sympathetic than you think. See if you can work together to restructure your job to make you more productive. From watching how you work, your supervisor may have some suggestions as to how you could do things more efficiently.
- **Delegate and learn how to say 'no'**
These two things are often linked to an unwillingness to pass on tasks — possibly because of a belief that if you want it done properly, you have to do it yourself. You may have to start trusting others a bit more.

**Certificate in Spoken and Written English III (Vocational English)
Assessment Competency 9: Can read an information text**

- **Aim to concentrate more while you work**
This way, it takes less effort to get more done. Set up your work environment so that you have peace and quiet. Try to confine conversations with other colleagues to lunch breaks.
- **Be more organised**
You should have everything you need at your fingertips.
- **Don't re-invent the wheel**
Be familiar with what other people in the office are doing (or have done) that could save you time.

**Certificate in Spoken and Written English III (Vocational English)
Assessment Competency 9: Can read an information text**

1. Number the topics below in the order they appear in the text.
 - Ways of dealing with pressure
 - The dangers of pressure
 - Types of pressure
 - The important things in your life
2. Crisis pressure and constant urgency are two types of pressure. Which one causes more health problems?

3. Many people overwork. What is the example given in the text of a heavy work program?

4. When people are overtired their work can be affected. Give two examples of this.
 - a. _____
 - b. _____
5. Who may be able to help organise your work to reduce stress and pressure?

6. Give two ways to improve your concentration while working.
 - a. _____
 - b. _____

**Certificate in Spoken and Written English III (Vocational English)
Assessment Competency 9: Can read an information text**

Answer key: Pressure

1. 4 = ways of dealing with pressure
2 = the dangers of pressure
1 = types of pressure
3 = the important things in your life
2. constant urgency
3. working 50 and 60 hours per week
4. productivity
decision-making skills
5. your supervisor
6. have peace and quiet
only talk to other colleagues in the lunch break

**Certificate in Spoken and Written English III (Vocational English)
Assessment Competency 9: Can read an information text**

Read the text then answer the questions following.

STUDYING TAFE COURSES with **OTEN** *the options*

WHAT IS OTEN?

The Open Training and Education Network (OTEN) is a joint initiative of the NSW TAFE Commission and the Department of School Education. OTEN offers over 200 TAFE NSW and secondary school courses to more than 28 000 students.

When you complete a TAFE NSW course through OTEN you receive the relevant TAFE NSW award.

OTEN offers you study options that are not usually available in other TAFE NSW Institutes. For example, you can learn at your own pace and, for most courses, you will study at home.



WHAT STUDY OPTIONS DOES OTEN OFFER?

Do I have to attend classes?

Most OTEN courses do not require attendance at classes. Learning materials that are specially designed to help you study without having to attend class are mailed to you. In these 'distance' courses, you still have access to a teacher by mail or telephone. There are also staff who offer extra support to students with special circumstances or needs.

Some OTEN courses have practical components such as

field trips, laboratory sessions or tutorials. In many cases you can study part of your course by distance education and part at another TAFE NSW campus where you attend classes.

When do I have to complete my course?

There are time limits for completing courses and subjects. It may be possible to complete your subject or course at a faster pace than you can in other TAFE

NSW Institutes. Alternatively you can work at a slower pace provided you complete your studies within the time limits.

Can I transfer to and from OTEN?

You can apply to transfer to OTEN from another TAFE NSW campus to complete part or all of your course providing you discuss this with your campus of original enrolment. You can also transfer from OTEN to another TAFE NSW campus.

WHAT TAFE NSW COURSES DOES OTEN OFFER?

OTEN offers a wide range of courses in these areas:

- building and construction
- business
- real estate
- office administration
- accounting
- engineering
- computers
- manufacturing
- hairdressing
- health
- languages
- rural studies
- tourism and hospitality
- maritime studies

OTEN also offers TAFE NSW courses such as:

- Certificate in General Education
- Matriculation
- English for Speakers of Other Languages
- Adult Basic Education
- Literacy and Numeracy.

A number of professional development courses on a fee-for-service (commercial) basis are also offered. These include management, building, business, driving instruction, liquor licensing and audiometry.

Contact a course information officer at OTEN or any TAFE NSW campus for a more detailed listing of courses offered by OTEN.



Reproduced by permission of OTEN

**Certificate in Spoken and Written English III (Vocational English)
Assessment Competency 9: Can read an information text**

1. What does 'OTEN' stand for?

2. OTEN offers hundreds of TAFE NSW courses. What other courses does it offer?

3. What are two special features of the courses offered through OTEN?
 - a. _____
 - b. _____
4. How do you keep in contact with teachers if you study through OTEN?

5. What practical learning activities may you be required to attend?

6. Where can you get more information about OTEN courses?

**Certificate in Spoken and Written English III (Vocational English)
Assessment Competency 9: Can read an information text****Answer key: OTEN**

1. Open Training and Education Network
2. secondary school courses
3. learn at own pace (at faster or slower pace)
study at home/don't have to attend classes
4. telephone or mail
5. field trips, laboratory sessions or tutorials
6. course information officer at OTEN or TAFE NSW

**Certificate in Spoken and Written English III (Vocational English)
Assessment Competency 9: Can read an information text**

Read the text then answer the questions following.

Migraine: Not just a headache

For those who don't suffer from migraines, it is almost impossible to understand how debilitating they can be. They can hit anywhere, at any time, and can last for a few hours to three days. The majority of people who get migraines are between 25 and 34 years of age. More women than men suffer from migraines.

A migraine headache has distinct symptoms which make it different from other headaches. Sometimes the pain is so severe that people may start to wonder if it more than just a migraine. And you can't be sure that it isn't until you visit your doctor who will eliminate other possibilities. You should never diagnose yourself.

Symptoms

- Moderate to severe pain, usually on one side of the head, that is pulsating or throbbing
- Nausea and possibly vomiting
- Sensitivity to light and sound (which sometimes becomes almost unbearable)
- Sometimes double vision and flashes of light

What triggers migraines?

Although there is no conclusive evidence that can be given as an explanation for all migraines, sufferers recognise that certain factors can trigger attacks.

Common triggers include:

- missing meals
- low blood sugar levels
- a deficiency in certain vitamins and minerals
- certain foods such as chocolate, cheese and citrus fruits
- alcohol
- too little or too much sleep
- certain smells
- stress or fatigue
- hormonal imbalances, menstrual cycle, or oral contraceptive use

Treatment

- Keep a 'migraine diary'. Note in it any factors such as the ones listed above that can trigger an attack. By looking back through your diary, you may be able to work out a pattern of triggers and then avoid them.
- Learn relaxation techniques such as stretching, yoga or massage.
- Try sleeping off a migraine.
- Try taking some physical exercise when you feel a migraine coming on.
- Note any sort of warning sign that a migraine is coming on. Once you know you are going to have one, you can take medication that should be prescribed by your doctor.
- Try a natural therapy such as herbs or vitamin and mineral supplements, but only try these after you've had your migraines diagnosed by a doctor.

**Certificate in Spoken and Written English III (Vocational English)
Assessment Competency 9: Can read an information text**

1. Who suffers most from migraines? Give all the details.

2. Why is it important to have your migraines diagnosed by a doctor?

3. What is one indicator that someone has a migraine?

4. Which foods are thought to cause migraine?

5. Why is it helpful to keep a 'migraine diary'?

6. At what time can physical exercise be helpful to migraine sufferers?

**Certificate in Spoken and Written English III (Vocational English)
Assessment Competency 9: Can read an information text**

Answer key: Migraine

1. more women than men, aged between 25 and 34
(Must have all above information)
2. it might be something else; may be more than just a migraine; eliminate possibilities of other problems
(Any one)
3. moderate to severe pain (usually on one side of the head); nausea and vomiting; sensitivity to light and sound; double vision and flashes of light
(Any one)
4. chocolate, cheese, citrus fruits
(Must have all three)
5. to work out a pattern of triggers/causes and then avoid them
6. when you feel a migraine coming on

Appendix 4

Examples of CSWE Oral Assessment Tasks (manipulations not included)

Certificate II tasks

Competency 5

Task 1

You invite a friend from your class to visit you. Give instructions on how to get from where you are now to where you live.

Tasks 2 and 3

Your friend has a new flexi card for an Automatic Teller Machine (ATM). Explain how to use the card to get money from the ATM.

Tasks 4 and 5

The globe in your kitchen light doesn't work. It is on a high ceiling. Give instructions to your 12 year-old child on how to change it.

Competency 6

Task 1

You want to visit Alice Springs for one week in September. Call the airline company and ask about a plane trip to Alice Springs.

Tasks 2 and 3

There is an exhibition of Aboriginal paintings at the museum which you would like to see. Call the museum and obtain information about the exhibition.

Task 4

You want information about English classes for yourself. Enquire at a local teaching centre.

Task 5

Your family has moved to a new suburb. You want information about the local high school for your 13 year-old child. Phone the school.

Competency 7

Task 1

Your washing machine doesn't work. Call a repair service and ask for someone to come and fix it.

Tasks 2 and 3

Your television doesn't work — you can't get the ABC. Call the television repair service.

Tasks 4 and 5

You want to have the Australian newspaper delivered to your home on Saturdays. Call your newsagent.

Certificate III tasks

Competency 5

Task 1

Ring the TAFE Information Centre and enquire about computer courses. Say what type of course you are interested in.

Tasks 2 and 5

Read the job advertisement below. Call the contact person and obtain information about the position.

(an advertisement was reproduced for the task)

Competency 6

Task 1

You have an appointment for a job interview with an employment agency tomorrow. The time that has been arranged is not convenient for you. Go to the agency, introduce yourself and explain the situation. Try and arrange another time for the interview.

Tasks 2 and 3

You have four weeks annual leave available this year. You would like to take three weeks leave now, even though it is a busy time at your workplace. Talk to your manager about this situation, explain why you want to take the leave now and negotiate a solution.

Tasks 4 and 5

You are an employee in a department store. A customer approaches you with a complaint about a faulty cassette recorder that they bought. The shop's policy is to only give credit for returned goods and not to give money back. Attend to the customer's complaint.

Appendix 5

CSWE Writing Assessment Tasks

**Certificate in Spoken and Written English III (Vocational English)
Assessment Competency 10: Can write a procedural text**

- Write a set of instructions on how to operate a cassette recorder.
Include information about:
 - playing cassettes
 - rewinding
 - fast forwarding
 - recording
 You can use your dictionary
Write about six instructions
- Write a set of instructions on how to enrol in an English course at AMES.
You can use your dictionary
Write about six instructions
- Write a set of instructions on how to apply for Australian Citizenship.
You may use the information provided by the Department of Immigration and Ethnic Affairs.
You can use your dictionary
Write about six instructions

Created by NCELTR as trialling material (1996).

**Certificate in Spoken and Written English III (Vocational English)
Assessment Competency 12: Can write a report**

- Write a report about a trade or profession in either your country of origin or Australia.
You can use your dictionary
Write about 150 words
Time: 1 hour
One of a bank of teachers' tasks sent to NCELTR Research for appraisal (1996).
- Write a report describing and comparing the four hotels in the table below and recommend one or more.
You can use your dictionary
Write about 150 words
Time: 1 hour

Name	Price of a room	Atmosphere and decor	Service	Breakfast
The Metropole	\$89.00	Elegant and luxurious	Rather leisurely	**** Excellent A wide choice of food
Hannan's Hotel	\$79.00	Modern, noisy	Efficient service	** Good, but unexciting
The Inn	\$65.00	Dull, shabby	Slow	* Disappointing Poor choice of food. Food cold.
The Blue Duck	\$47.00	Lively, cheerful atmosphere	Good service, friendly staff	*** Delicious home cooking

This table has been reprinted from the draft, 'Competency Based Assessment Tasks for the CSWE Further Study' compiled by Nita Johnson, AMES WA.

Certificate in Spoken and Written English III (Vocational English) Assessment Competency 12: Can write a report

3. Your manager wants to purchase a new frost free refrigerator for the staff room. The old one is much too small and has no freezing compartment. She wants a quality product but doesn't want to pay a high price.

She has asked you to write a report outlining several options. She would also value your opinion as to the best purchases.

You may use the information from *Choice* magazine for your information.¹

You can use your dictionary

Write about 150 words

Time: 1 hour

¹ The input text for this task was taken from *Choice* magazine. For copyright reasons the text cannot be reproduced.

Created as a trialling task modified from an AMES-CSWE benchmark task by NCELTR (1996).

Notes on contributors

Geoff Brindley is a senior lecturer in the Department of Linguistics at Macquarie University and Research Coordinator in the National Centre for English Language Teaching and Research. He is the author of a range of publications on language proficiency assessment, second language acquisition and curriculum design.

David Smith BA, Dip Ed. has a Postgraduate Diploma in Educational Studies (TESOL/ALBE) and a Master of Education Degree from the University of Melbourne. His Masters thesis explored the issue of rater consistency and judgement in the assessment of written competencies within the Certificates in Spoken and Written English. David has worked for several years as an adult ESL teacher for AMES. During this time he developed and delivered a range of workplace language training programs. He is currently employed as the Learning and Development Coordinator for a large multinational organisation.

Gillian Wigglesworth is a senior lecturer in Linguistics and Convenor of the Master of Applied Linguistics programs at Macquarie University. Her research interests encompass first language acquisition, second language acquisition and the evaluation and assessment of second languages. She has published both in Australia and internationally on issues related to language assessment, second language acquisition and linguistic and cognitive development of school age children. She has coedited three books, one on language and gender, and two on issues in language testing and evaluation.

Steven Ross is Professor of Linguistics in the School of Policy Studies, Kwansai Gakuin University, Japan. He worked as a research associate at the National Centre for English Language Teaching and Research, Macquarie University in 1994. His current research is in second language acquisition, language assessment, and program evaluation.