This volume brings together several related research studies concerning assessment practices in the Adult Migrant English Program (AMEP). The studies focus on the competency-based achievement assessment system which accompanies the Certificate of Spoken and Written English (CSWE), the outcomes-based curriculum framework used nationally in the AMEP. Part one of this volume addresses issues of consistency of the assessment procedures and practices used in the AMEP. Part two outlines the results of a large-scale study into the ‘washback’ effect of the implementation of CSWE assessment.

Studies in immigrant English language assessment: Volume 2 is a companion to Studies in immigrant English language assessment: Volume 1. It will be of interest to teachers, curriculum developers, researchers, government analysts and officials, and decision makers with an interest in language performance assessment and its interface with language policy.

Studies in immigrant English language assessment
Volume 2
Edited by Geoff Brindley and Catherine Burrows
Studies in immigrant English language assessment
Volume 2
Edited by Geoff Brindley and Catherine Burrows

RESEARCH SERIES 11 2001
National Centre for English Language Teaching and Research
Macquarie University Sydney NSW 2109
Contents

Figures v
Tables v
Acronyms viii
Introduction 1

Part 1
Chapter 1 15
Assessment and moderation in the CSWE: Processes, performances and tasks
Catherine Burrows

Chapter 2 59
Investigating rater consistency in competency-based language assessment
Stephanie Claire

Appendix 1 81
Text suitability checklist

Part 2
Chapter 3 95
Searching for washback: The impact of assessment in the Certificate in Spoken and Written English
Catherine Burrows

Appendix 1 185
Interview schedule
List of Figures and Tables

Figures
Figure 1: All teachers: Lessons Content combined % 151
Figure 2: All teachers: Sub-categories of Form only, lessons combined % 152
Figure 3: All teachers: Both lessons, minimal, extended and minimal + extended texts % of total lesson time 155
Figure 4.1: Traditional washback theory: A stimulus-response model 169
Figure 4.2: 1990s view of washback: A ‘black box’ model 169
Figure 4.3: Proposed view of washback: A curriculum innovation model 170

Tables
Table 1: Preferred assessment methods 17
Table 2: Assessment of oral competencies: Other teachers involved 20
Table 3: Audiotaping of student performance 21
Table 4: Frequency of audiotaping 21
Table 5: Comments on the process of audiotaping students 22
Table 6: Videotaping of student performance 23
Table 7: Frequency of videotaping 23
Table 8: Comments on videotaping assessment performances 24
Table 9: Moderation sites and competencies assessed 28
Table 10: Certificate 1 Competency 4: Can provide personally relevant information using spoken language 63
Table 11: Certificate 2 Competency 5: Can give spoken instructions 64
Table 12: Certificate 3 Competency 6: Can negotiate a complex/problematic spoken exchange 65
Acronyms

AMEP  Adult Migrant English Program
AMESs  Adult Migrant English Service
AMES  Adult Multicultural Education Services
AR    Assessment and Referral
ASLPR Australian Second Language Proficiency Ratings
CBT   Competency-based training
COLT  Communicative Orientation of Language Teaching
CSWE  Certificates of Spoken and Written English
DIMA  Department of Immigration and Multicultural Affairs
ESB   English speaking background
ESL   English as a Second Language
IELTS International English Language Testing System
NCELTR National Centre for English Language Teaching and Research
NESB  non-English speaking background
pc    performance criteria
TOEFL the Test of English as a Foreign Language
Introduction

Catherine Burrows

This volume, a companion to Volume 1 (Brindley 2000a), brings together three studies of assessment practices in the Adult Migrant English Program (AMEP) in Australia. These studies all focus on the competency-based achievement assessment system which accompanies the Certificates in Spoken and Written English (CSWE), the national curriculum framework that is used in the AMEP. The first two studies by Brindley and Claire in Part 1 were conducted in 1997–98 as part of the Special Projects Research Program, conducted by the National Centre for English Language Teaching and Research (NCELTR) and funded by the Department of Immigration and Multicultural Affairs (DIMA). These studies used different but complementary methodologies to examine the consistency of the assessment procedures and practices used in the AMEP. The third study, reported in Part 2 by Burrows, was conducted as doctoral research at Macquarie University and examined the washback effects of the implementation of CSWE assessment.

The aim of this chapter is to situate these studies within the social, political and educational context of the AMEP and to identify some of the key common issues that emerge. For more extensive description and analysis of the assessment procedures used in the AMEP, the reader is referred to Brindley (2000a).

Background: Teaching and learning in the AMEP

In Australia, English language and literacy teaching to adult immigrants is provided through the DIMA-funded AMEP. The AMEP was founded as part of a plan to deal with ‘... problems associated with the “assimilation of foreigners” ... when non-English speaking immigrants began to settle in numbers in Australia after the end of the Second World War’ (Armstrong 1969:7, in Campbell 1986:3).

The plan of the AMEP was to provide instruction in English language and settlement information to non-English speaking migrants at four stages: pre-embarkation; on board ships coming to Australia; on-arrival; and after placement in employment. ... throughout its history the AMEP has been primarily an element within the post-arrival services for migrants. (Campbell 1986:3)
The AMEP is delivered by a number of educational providers in each state, known until 1998 as the Adult Migrant English Services (AMESs). Since that time the AMEP has been made available to immigrants and refugees through a variety of public and private sector agencies.

During the period following the establishment of the AMEP, a wide range of different teaching approaches, methods, styles and syllabuses had been used by teachers within the AMESs, some concurrently, ranging from the notional–functional syllabus to ‘unconventional’ approaches such as the Silent Way and Suggestopedia. Many teachers employed a needs-based approach, whereby they examined the needs of each particular group of students and attempted to design a syllabus specifically tailored to address those needs. Not all students, however, found this approach satisfactory:

As part of the Campbell Review of the AMEP (1985) learners were surveyed and ‘the absence of a coherent progression of courses’ was identified as an obstacle to learner achievement … Teachers also identified lack of continuity as a problem … Although committed to a learner-centred curriculum, many teachers expressed difficulty in implementing the needs-based approach and pointed to the need for more centralised curriculum planning. (Hagan 1994:30–1)

**Learner pathways**

The first direct result of these concerns was the implementation of learner pathways (Colman 1990, 1991; Burton 1991; Hagan 1994) which were designed to give sequence and structure to courses and, in particular, from course to course. In this context, Hood (1995:22) described the CSWE as … the outcome of a process of change which began in the late 1980s with the publication of a series of guidelines on course design … By 1992, the process of developing a more systematic approach to curriculum was well underway with a national Learner Pathways initiative.

NSW AMES also introduced the concept of Bands, or pace of learning, which were to become important as DIMA acknowledged that different types of students would take different amounts of time to reach the same levels.

Band A indicates a slower paced pathway usually for students with 0–6 years education. Band B, standard pace for students with 7–12 years education, and Band C, fast pace for students with tertiary education.

These definitions are not rigid and students can move across bands if their original placement was not appropriate. (Hagan 1994:5)

When students enter the Program, they are placed in a class according to both Level and Band: a Band A student, for example, is not expected to learn as quickly as a Band C student. This is beneficial to both student and educational provider, as slower-paced students may be placed in a class with fewer hours per week to allow for adjustment to a new educational setting, while faster-paced students can progress more rapidly through the Program, if they choose, by attending higher intensity provision.

**Competency-based assessment in the AMEP: The CSWE**

In 1993, a new curriculum framework, the Certificate in Spoken and Written English (CSWE) (Hagan et al 1992), was introduced into the AMEP. The CSWE and its assessment are described in Brindley (2000b), along with a discussion of various issues and problems associated with its development and use. For this reason, the following description will be brief.

Hood (1995:22) describes the CSWE as ‘… competency-based curriculum framework structured around a social and functional theory of language …’. Its structure and organisation are characterised by Sturgess (1996:21) as follows:

The accredited and nationally registered CSWE sets out a framework for consistently describing what learners can do in English as a result of the tuition they have received. It consists of three stages — beginners (ASLPR 0–0+), post-beginners (ASLPR 1– to 1) and intermediate (ASLPR 1+ to 2) and recognises three different learning paces based on the student’s previous learning experience — slow, standard and fast. In Stage 3 there is provision from three different specific learning focuses, depending on the individual student’s aspirations — English for study, vocational English and community access.

The current version (NSW AMES 1998) includes four Certificate levels, each with a series of competencies covering study skills, speaking, listening, writing and an optional numeracy module. Teachers design syllabuses to meet the needs of their students within the framework provided by the curriculum. Students are assessed against combinations of competencies which allow the achievement of Certificates or of modules within Certificate Levels (e.g. literacy; or speaking and listening).
Since its implementation, the CSWE has been progressively updated in a number of editions (NSW AM ES 1995; NSW AM ES 1998), as required by accreditation standards bodies. It has been amended according to feedback from stakeholders, including teachers and educational managers, as well as feedback from students. Modifications have also been carried out in the light of the outcomes of research (eg Brindley 2000c, 2000d; Wigglesworth 2000) and statistical analysis conducted by DIMA's National Management Information Unit. Alterations of a different order, for example the shift from one Certificate to four, were required by changes in governmental procedures related to accreditation of curriculum.

The role of assessment in the CSWE
With the CSWE came competency-based achievement assessment, which was situated wholly within the classroom. In order to determine Certificate achievement, assessment tasks are designed and administered by teachers, using model tasks created by the curriculum designers. Before the implementation of the CSWE, teachers reported the outcomes of student learning against a broad-banded general proficiency rating scale, the Australian Second Language Proficiency Ratings (ASLPR) (Ingram and Wylie 1984), which was also used to place students initially. At this time, assessment practices were not consistent across Australia: some states reported on proficiency measures for all macro-skills while others reported only on spoken language. This form of reporting was apparently seen to be sufficient, especially since students were placed in classes largely according to their spoken language proficiency.

Subsequently, along with the added emphasis on program accountability that characterised the 1990s, AM EP outcomes began to come under closer scrutiny from government authorities. Examination of aggregated program outcomes in terms of ASLPR gains revealed that only 17% of students were reaching the stated goal of ASLPR 2, or 'Functional English' proficiency. To teachers, these results were not surprising since most of the learners in the AM EP are below the level of 'survival proficiency' as defined by the ASLPR and a 'broad brush' general proficiency scale such as the ASLPR is not sensitive enough to register the gains in specific competencies that such learners make (Cope et al 1994; Brindley 2000b). However, these ostensibly modest outcomes put the Program at risk of losing its Federal support and therefore its funding. It was clear that the DIMA needed a way to show progress that would accurately show the gains made by all learners, however small. At the same time, the effects of individual differences on learning outcomes needed to be acknowledged. In particular, it was clear that it would take AM EP clients with limited education and low levels of language proficiency a considerable period of time to reach functional proficiency.

Two steps were taken to remedy this problem. The first was that the Immigration Minister of the day, Minister Hand, put into place legislation which ensured that new immigrants whose English is below the stipulated level of ASLPR 2, received an entitlement of 510 hours of tuition. This eligibility ceases once an immigrant reaches ASLPR 2 or uses 510 hours of tuition. Later, achievement of Certificate III from the CSWE was added as equivalent to ASLPR 2.

The second step taken to ensure a more accurate recording of progress was the development of the CSWE, by the New South Wales Adult Migrant English Service (NSW AMES 1995, 1998). Since it is designed to reflect the achievement of particular tasks rather than to describe a generalised ability level, the CSWE has much smaller gradations than the ASLPR and allows teachers and educational providers to report students' actual achievement against prespecified outcome measures. This had a number of advantages for DIMA since it permitted much finer-grained and transparent reporting to funding authorities.

In the words of one of the key departmental decision-makers at the time:

One of the things they'd always been concerned about is very low level students with minimal education not making any gains on the ASLPR. And so what they got from this data was how you could show progress, and you actually were talking about it in ways that made sense to people ... and they could look at aggregating data which would actually show that the students had achieved certain competencies, modules, certificates, levels and so on. And that data could be broken down into learner profiles, years of education, language background, age and so on, all these reports show different ways in which students were progressing. (Cited in Moore 2001:186)

The assessment of the CSWE requires that students achieve competencies in the different language skill areas specified above. The first edition also included competencies concerning knowledge of the Australian context but these were removed from the third edition, following negative feedback from teachers relating to the difficulty of making accurate judgments on these, as well as the results of a study conducted by Aldred and Claire (1994) which raised concerns about the reliability of the assessment of these competencies.
Research into the CSWE

The competency-based model that underpins the CSWE has been widely criticised by researchers who do not support competency-based training per se (see, for example: Kenway, in Reid 1992; Walker 1993; Toms 1995) or do not believe that the adaptation of competency-based training and its assessment to language learning is appropriate (see, for example: Quinn 1993; Grove 1996; McNamara 1999). Partly as a result of these criticisms, the CSWE has been the subject of a series of research projects, many concerned with investigating the validity and reliability of CSWE assessment. These have included various studies by Brindley (eg 1994; 2000b, 2000c, 2000d), Wigglesworth (2000), Burrows (1993), Jones (1993), Aldred and Claire (1994) as well as the studies reported in the present volume. Other studies concerned with the CSWE have included the work of Ross (2000) which has been concerned with modelling the impact of individual differences on CSWE outcomes; studies by Burrows (1998; this volume) which focused on the washback effects of CSWE assessment, and the work of Burns, Hood and Joyce (1995, 1997, 1998 and 2000) which has involved a series of action research projects conducted by teachers and related to various aspects of teaching and assessing the CSWE. The CSWE is thus the object of an ongoing research and evaluation program, the results of which have been used to modify the content in successive versions.

Assessment and reporting: The current climate

In 1997, DIMA put the Adult Migrant English Program out to public tender. Many, but not all, of the previous providers of the AMEP remained within the Program, although some with a considerably smaller share of the provision. With the tendering of the AMEP, an element of open competition was introduced between providers and, consequently, accuracy regarding assessment and outcomes reporting has become critically significant. In a climate of greater public accountability for government expenditure, providers of the CSWE must be able to demonstrate to DIMA the reliability of their assessment results and reporting procedures. DIMA, in order to make policy decisions that are equitable to all stakeholders, is reliant on receiving accurate and reliable CSWE data from all AMEP providers. In turn, the performance of AMEP providers is measured against the outcomes of their students, using the ‘benchmarks’ as described above. This, too, increases the importance of reliable assessment in the AMEP.

The current volume

The research reported in the first volume ‘spans a wide range of assessment concerns, ranging from issues of macro-level reporting through to the function of individual performance criteria’ (Brindley 2000c:34). The studies reported in this volume cover a somewhat narrower range of assessment concerns, each focusing in a different way on classroom-based assessment.

The first chapter by Claire reports on a qualitative study of moderation and assessment practices in the CSWE. Claire investigates the conduct of CSWE moderation exercises and examines rating practices for CSWE assessment. She looks at the way in which judgments are reached during moderation exercises and particularly at the manner in which raters ‘agree to disagree’. She also analyses the moderation of spoken language competencies and investigates the extent to which student spoken language is recorded for assessment purposes. Claire identifies a number of practices current at the time of her study which affect consistency in assessment and makes a series of recommendations as to how assessment and moderation might be improved.

In Chapter 2, Brindley addresses the question of consistency of text classification by CSWE assessors. Based on an analysis of the rating patterns of 12 experienced raters, he finds that trained assessors do not consistently agree on the classification of texts submitted by teaching centres as ‘benchmark’ examples of minimal competency achievement. He also identifies marked differences in the interpretation of performance criteria and that the criteria appeared to be of a different order of difficulty, findings which have also emerged from previous studies of CSWE assessment (Brindley 2000d; Smith 2000). According to Brindley, the low levels of consistency in classification of texts which were identified suggest 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. The author also highlights the need for professional development and assessment task banking to improve consistency in assessment, echoing the concerns expressed by Claire in Chapter 1.

In Part Two of this volume, Burrows looks at the question of the washback of CSWE assessment. She examines the impact of the introduction of CSWE assessment on teachers and their teaching practices, through the use of a survey, interviews and structured analysis of classroom observations. The main finding is a divergence between the responses of different teachers, which would appear to be dependent on their beliefs, assumptions and knowledge.
Finally, in an era of economic rationalism, when educational institutions are by necessity keeping expenditure as low as possible in order to compete successfully for funding, the importance of maintaining professionalism in language education is paramount. Continued professional development and assessment materials development are essential for the AMEP to continue to support its clients with high quality educational programs and for DIMA to continue to report successful outcomes from the Program it funds.

Endnotes
1 In 2000, the role of NCELTR in servicing the research and development needs of the AMEP was taken over by the AMEP Research Centre, a consortium involving Macquarie University through NCELTR and La Trobe University through the Institute for Education.
2 Brindley (2000c), however, questions the extent to which this equivalence can be empirically supported.

References


English language assessment, vol 1. Sydney: NCELTR, Macquarie University


Burton, J. 1991. 'Perceptions of learning in the AMEP'. Prospect, 7, 1: 57-72


Colman, J. 1991. 'Towards a coherent curriculum frame: Learner pathways in the NSW Adult Migrant English Service'. Prospect, 7, 1: 28-42


McNamara, T. 1999. July. 'Validity in language testing: The challenge of Sam Messick’s legacy.' Plenary address presented at 21st Language Testing Research Colloquium. Tsukuba, Japan


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

Reid, A. (ed) 1992. 'Point and counterpoint: Accountability and education: The great profile debate'. Curriculum Perspectives, 12, 1: 55–78


PART 1
1
Assessment and moderation in the CSWE: Processes, performances and tasks
Stephanie Claire

Introduction
This chapter reports the results of two related studies into the assessment and moderation practices that are used in conjunction with the Certificates of Spoken and Written English (CSWE). The first part of the chapter outlines the results of a survey of CSWE assessment processes and discusses implications of the findings for assessment policy and further research. The second part of the chapter reports the findings of a qualitative study of moderation activities in a number of different AMEP centres and evaluates the extent to which the practices in place were found to be effective. The chapter concludes with recommendations for ways of improving assessment and moderation practices.

Background
The studies reported in this chapter arose from the need to ensure that assessment and reporting within the CSWE was sufficiently rigorous to support its increasingly high stakes. Although originally intended to facilitate teachers’ assessment of student progress and achievement, the CSWE and its assessment have been progressively adopted as both an assessment and reporting mechanism at a system-wide level. As the mandatory curriculum for the AMEP, the CSWE is used by adult migrant education providers, Technical and Further Education and Adult and Community Education agencies. It has also been adopted in school education contexts in New South Wales and in Aboriginal Education in the Northern Territory of Australia. Outside Australia, it has been extensively drawn on by curriculum and assessment projects in South Africa, Mauritius and Canada and was recently adopted as the curriculum for adult English language teaching in East Timor by a number of agencies.
As well as its growing high public profile, the stakes associated with CSWE assessment have increased. In 1996, CSWE competency outcomes began to be used as a basis for deriving ‘benchmarks’ which reported average student achievement of competencies in terms of instructional hours (Ross 2000). These benchmarks have been used by DIMA as a means of monitoring and evaluating provider performance, even though the original purpose of CSWE assessment was primarily to inform teachers and learners (Burrows 1993). Information derived from teacher-based assessments is thus assuming considerable importance as an accountability measure and may in future be used to determine ESL learners’ access to opportunities such as further education, retraining, work and even to promotion. For this reason, CSWE assessments need to be demonstrably valid and consistent.

**Study One: Classroom-based assessment processes**

**The survey**

As a prelude to the review of moderation practices, a small survey was conducted in order to investigate the ways in which teachers were conducting assessments within their own organisations. The survey was designed to provide information on:

- the degree of collegial involvement in the assessment process
- the actual process used by teachers in making formal assessment judgments
- the extent to which teachers record student performances on audiotape or videotape.

A total of 43 AMEP teachers from three Australian states (New South Wales, Victoria and Western Australia) were surveyed. Of these 15 were from NSW AMES, 9 were from AMES Victoria, and 19 were from AMES Western Australia. The participants from NSW AMES were CSWE representatives, that is, teachers who had volunteered to liaise with the NSW AMES Program Support and Development Services unit on CSWE-related matters. During 1997 the CSWE representatives had attended a ‘benchmarking workshop’ at which they moderated written texts that had been previously identified by regions as ‘benchmarks’, that is, texts exemplifying minimal achievement of all performance criteria. The Victorian and Western Australian participants were AMES teachers who were not identified as being part of a special CSWE ‘interest group’.

### Survey results

The first question in Part 1 of the survey asked respondents to identify from a number of options the assessment methods they felt were best for their students. They were also given the opportunity to nominate other methods. Responses are summarised below in Table 1. In interpreting this table, it should be noted that some teachers indicated more than one option as their preferred method of assessment.

<table>
<thead>
<tr>
<th>Personnel involved in assessment</th>
<th>Total respondents choosing this method</th>
<th>Respondents choosing this method by state</th>
</tr>
</thead>
<tbody>
<tr>
<td>All students assessed by class teacher</td>
<td>18</td>
<td>NSW 6 5 7</td>
</tr>
<tr>
<td>All students assessed by another teacher</td>
<td>5</td>
<td>Vic 4 0 1</td>
</tr>
<tr>
<td>Consultation with another teacher on borderline/problematic texts</td>
<td>37</td>
<td>WA 13 11 13</td>
</tr>
<tr>
<td>Other</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>
It can be seen that while the respondents primarily assessed their own students in class (a total of 18 respondents in comparison with five respondents who reported that all their students were assessed by another teacher), they also engaged in a considerable amount of collegial consultation regarding ‘difficult’ texts: a total of 37 teachers reported that they consult other teachers on borderline or problematic texts.

Other approaches that were nominated included:

- the use of different methods for different Certificate levels: the first option (all students assessed by class teacher) for Certificate I, the second option (all students assessed by another teacher) for Certificates II and III and the third option (consultation with another teacher on borderline/problematic texts) for use at all levels
- the use of Assessment and Referral (AR) teachers in the assessment of spoken and listening competencies (AR teachers are teachers specifically trained to undertake placement assessment).

Fifteen respondents (NSW = 12, Victoria = 0, WA = 3) chose to comment. They mentioned the problems of logistics (such as time, space, availability of other teachers) regarding the involvement of colleagues in assessment, and the particular importance of collegial involvement in the assessment of spoken competencies. One warned of the danger of bias towards ‘nice students’ and another argued that ‘the teacher’s awareness of the particular student’s “body of work” over a term or two is the best indicator of a pass’. One respondent advocated the assessment of students by their classmates, especially in Level III.

The first question in Section 2 of the survey asked respondents to outline the process they used to assess student performances of oral competencies. Forty of the 43 respondents (NSW = 12, Vic = 9, WA = 19) answered this question, of whom 22 provided a clear description of the processes they used. Examples included the following:

- Decide on task, check task meets range statements and performance criteria.
- Often students practise a similar task and either self- or peer-assess prior to the test being given.
- Task given, discuss what is required.
- Feedback given to students. Students (who) are not assessed as having achieved (are) given later opportunity to try again.

- Assess whether student has fulfilled the performance criteria and demonstrated competence as required in the guidelines.
- Has the student shown required ability in this competency? Have the elements been fulfilled? Have the performance criteria been demonstrated?

The remainder of the responses were less precise. Although the assessment activities described may have been consistent with CSWE requirements, it was unclear what form these activities took. This imprecision is illustrated in comments below.

- When I feel the student has had adequate preparation I withdraw the student to perform the task. Sometimes the competency requires a listener other than the teacher. For these I ask another teacher to assess. For borderline students, I ask another teacher to assess.
- Depends on the level. Could be observation of group, could be one-to-one with teacher.
- Students withdrawn from class by teacher individually to be assessed, with some oral assessment.

Other teachers mentioned the role of observation in their assessment process, but in some cases it was not clear exactly what was being observed or how the performance criteria were being applied:

- (I use) ongoing observation when engaging student in conversation or when student is speaking to others.
- In Cert. I/II, observation, chatting. Cert. III, perhaps a move in conversation. I don’t tell students I’m assessing them on conversation, because it’s so unnatural.

Overall, the responses suggested that the teachers who were surveyed conformed to the CSWE assessment requirements, with 22 of them mentioning one or more of the following terms: competency, performance criterion, checklist, element, range statement in their replies. However, responses such as ongoing observation when engaging student in conversation or when student is speaking to others raise doubts concerning the rigour of the assessment.

The second question in Section 2 of the survey asked respondents to identify the degree of involvement on the part of other teachers in their assessment of oral competencies. Responses are shown in Table 2.
Table 2: Assessment of oral competencies: Other teachers involved

<table>
<thead>
<tr>
<th>State</th>
<th>Yes</th>
<th>No</th>
<th>Total Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSW</td>
<td>12</td>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td>Vic</td>
<td>6</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>WA</td>
<td>12</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>Total</td>
<td>30</td>
<td>10</td>
<td>40/44</td>
</tr>
</tbody>
</table>

The responses to this question indicated a high degree of involvement by other teachers in the assessment of spoken language. Overall, 30 teachers reported collegial involvement in assessment compared with a total of 10 who reported no such involvement. Of these respondents, 28 described the use of other teachers (NSW = 9, Vic = 7, WA = 12).

The ‘other teachers’ identified in the responses included ILC teachers, coordinators and ‘roving assessors’. Non-teachers such as volunteers, vocational counsellors and public officials were also mentioned, although non-teachers are not accredited to make assessment judgments.

The responses mentioned a number of ways in which teachers or external personnel could be used in assessing students; some are shown below.

- Competencies which involve telephoning, e.g., I usually enlist the help of a colleague or assist another teacher by answering the phone.
- For competencies which involve negotiation skills, another teacher assesses students out of the class.
- Sometimes oral testing involved other teachers. Occasionally assessment is done when students interact with ‘non-teachers’ in a specific setting, e.g., asking for information while on excursions.

However, some responses to this question raised doubts as to whether the conditions of the assessment had been met. It is difficult to know, for example, how strictly performance criteria have been applied in cases where assessment is carried out by ‘asking other teachers to have a chat on excursions’.

The third survey question in Section 2 sought to establish whether respondents audiotaped student performances of oral competencies and, if so, how often they did this. This question elicited a total of 41 responses. The breakdown of responses by state is shown in Table 3.

Table 3: Audiotaping of student performance

<table>
<thead>
<tr>
<th>State</th>
<th>Yes</th>
<th>No</th>
<th>Total Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSW</td>
<td>11</td>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>Vic</td>
<td>5</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>WA</td>
<td>3</td>
<td>14</td>
<td>17</td>
</tr>
<tr>
<td>Total</td>
<td>19</td>
<td>22</td>
<td>41/44</td>
</tr>
</tbody>
</table>

Table 3 shows that slightly fewer than half of the teachers surveyed audiotaped student performances. This may indicate a wide variation between states regarding the practice of audiotaping student performances, with the teachers from NSW being much more likely to audiotape performances than teachers from the other states surveyed (NSW = 11/15, Vic = 5/9, WA = 3/19). In interpreting these results, however, it should be borne in mind, as noted above, that the sample of teachers is very small and the teachers from NSW AMES, the volunteer ‘CSWE representatives’, may not represent a typical cross-section of teachers in the Service.

The 19 respondents who had indicated that they audiotaped their students were asked to indicate how often they did this, using a frequency scale ranging from ‘always’ to ‘never’. In retrospect, however, it is clear that ‘never’ was not an appropriate option and none selected it. As shown in Table 4 below, ‘occasionally’ was the most frequent response, followed by ‘around 50% of the time’.

Table 4: Frequency of audiotaping

<table>
<thead>
<tr>
<th>Frequency of audiotaping</th>
<th>Always</th>
<th>Often</th>
<th>Around 50% of the time</th>
<th>Occasionally</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSW</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>Vic</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>WA</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>11</td>
<td>n = 19</td>
</tr>
</tbody>
</table>
Of those responding, 24 chose to comment on this question. Of these, 7 could be described as supporting the practice, or ‘favourable’, 11 could be described as negative towards the process, or ‘unfavourable’ and 6 could be described as ‘neutral’, although these comments were also informative. The breakdown by state was as follows:

<table>
<thead>
<tr>
<th>State</th>
<th>Favourable</th>
<th>Unfavourable</th>
<th>Neutral</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSW</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>13</td>
</tr>
<tr>
<td>Vic</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>WA</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>7</td>
<td>11</td>
<td>6</td>
<td>n = 24</td>
</tr>
</tbody>
</table>

‘Favourable’ comments included:
- Absolutely essential for students to hear their performance and refine it. Also important for pronunciation.
- When class numbers are not so high that this exercise is not feasible, this is useful to both students and teachers.

‘Unfavourable’ comments included:
- No time — difficulty in arranging equipment.
- It’s too time-consuming and I think it makes students even more nervous.

‘Neutral’ but informative comments included:
- Mainly group conversations.
- When the task involves real life situation — eg telephone the Council for information; interview or casual conversation with one or more English speakers.

Four teachers reported student performances ‘always’ or ‘often’ in comparison with the 15 teachers who reported audiotaping ‘around 50% of the time’ or ‘occasionally’. Perhaps not surprisingly, the preponderance of ‘unfavourable’ comments came from the latter group and may reflect a need on the part of these teachers to account for their low incidence of audiotaping.

There were 36 responses to the question, ‘Do you videotape student performances of oral competencies?’ The breakdown by state is shown in Table 6.

<table>
<thead>
<tr>
<th>State</th>
<th>Yes</th>
<th>No</th>
<th>Total Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSW</td>
<td>3</td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td>Vic</td>
<td>2</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>WA</td>
<td>2</td>
<td>14</td>
<td>16</td>
</tr>
<tr>
<td>Total</td>
<td>7</td>
<td>29</td>
<td>n = 36</td>
</tr>
</tbody>
</table>

Only 7 teachers replied in the affirmative, with 29 reporting that they did not videotape student performances, indicating a lower usage of video-recording than audio-recording.

There were 14 responses to the question, ‘If yes, how often do you videotape?’ As shown in Table 7 below, with one exception (NSW), ‘occasionally’ was the only response.

<table>
<thead>
<tr>
<th>State</th>
<th>Always</th>
<th>Often</th>
<th>Around 50% of the time</th>
<th>Occasionally</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSW</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Vic</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>WA</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>13</td>
<td>14</td>
</tr>
</tbody>
</table>

Respondents were invited to add their comments regarding the videotaping of student performances. Of the 25 comments received, 9 could be described as supporting the process, or ‘favourable’, 7 could be described as negative towards the process, or ‘unfavourable’, and 9 could be described as ‘neutral’ but informative. The breakdown by states was as shown in the following table.
which were perceived to be difficult to assess or 'borderline' were assessed informally by more than one teacher. The survey also revealed a degree of 'official' involvement of additional teachers to assist in assessment, such as teachers with specific responsibility for assessment and referral, although this was not consistent across respondents and, therefore, one might hypothesise, across providers.

However, the rather unsystematic involvement of additional teachers (mainly in cases perceived to be 'borderline') is a potential source of concern since the use of more than one rater is an important element in ensuring reliability. In this regard, the results of a number of research studies suggest that single ratings are quite unreliable, particularly when the number of assessment tasks is small (Brindley 2000b; McNamara and Lynch 1997).

The relative infrequency of audio and video recording of spoken performances is another factor affecting the quality of CSWE assessments. If no recorded sample of language is available, teachers are obliged to assess a number of performance criteria simultaneously on the basis of a live performance, a challenging task, even for the most experienced assessor. Under circumstances where only one hearing is involved and individual performance criteria cannot be checked exhaustively, there is a natural tendency to make a holistic judgment based on overall task fulfilment (McNamara 1996). This could affect not only the reliability of CSWE competency assessments but also their validity, since one could not be sure that the specified elements of a given competency were actually being assessed. On the other hand, if examples of spoken performances are recorded, they can then be used for moderation purposes, to the benefit of other teachers and the Program generally.

References by respondents to lack of time, equipment and resources may explain why recording does not occur more frequently, at least amongst this group of teachers. However, without recorded examples, there is a risk of compromising the quality of CSWE assessments but also their validity, since one could not be sure that the specified elements of a given competency were actually being assessed. On the other hand, if examples of spoken performances are recorded, they can then be used for moderation purposes, to the benefit of other teachers and the Program generally.

The results of the survey provided a useful picture of assessment processes used by teachers in the AMEP. While it is important not to generalise from a sample such as this one, the collated responses to the survey highlighted certain trends within the AMEP at the time of the survey. The first of these is that collegial involvement in assessment appears to have been widespread within the AMEP. This means that, at least in the case of the teachers surveyed, student texts

<table>
<thead>
<tr>
<th>Table 8: Comments on videotaping assessment performances</th>
</tr>
</thead>
<tbody>
<tr>
<td>State</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>NSW</td>
</tr>
<tr>
<td>Vic</td>
</tr>
<tr>
<td>WA</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

'Favourable' comments included:
- This particularly helped with interviews and short talks.
- Especially [useful] for 'negotiation' competency.

'Unfavourable' comments included:
- No support teacher available. Equipment heavy and obsolete. Stresses most students.
- Self-consciousness makes performance worse.
- Time-consuming and possibly intimidating.

'Neutral' but informative comments included:
- Videotaping of competencies is only done when students have been made accustomed to being filmed by regular use of the videotape in normal classroom teaching sessions.
- I videotape the students' performances [as] an additional teaching/learning tool, but not during the assessment performance.

As with the responses to the subsidiary question regarding frequency of audio-taping above, the large number of 'unfavourable' comments in comparison with 'favourable' ones may again be a rationale for not videotaping student performances but may also indicate a strongly held belief that this process is not practical.

Discussion and implications for assessment practice
The results of the survey provided a useful picture of assessment processes used by teachers in the AMEP. While it is important not to generalise from a sample such as this one, the collated responses to the survey highlighted certain trends within the AMEP at the time of the survey. The first of these is that collegial involvement in assessment appears to have been widespread within the AMEP. This means that, at least in the case of the teachers surveyed, student texts
issue it would be useful to investigate rating designs that enable economical means of double-rating student performance (for example, ‘minimal effort’ judging plans that do not require each performance to be rated by each rater [Linacre 1994]). This would not only enhance the consistency of teachers’ judgments but would allow rater performance to be monitored over time.

**Study Two: An examination of moderation practices for the Certificates in Spoken and Written English**

**Background**

Moderation is defined by Radnor and Shaw (1995:124) as ‘the review and social ratification of teachers’ assessments, that is, their judgements of the value of pupils’ work’. In the context of the CSWE, Christie and Delaruelle (1997:96) state that moderation is carried out ‘to ensure consistency in the application of standards within a teaching centre, across centres and across providers’.

As most assessment of student performance for the CSWE is carried out by individual teachers, with second rating largely reserved for borderline cases, as described above, moderation sessions involving teachers and program managers are used to support and maintain reliability (inter- and intra-teacher and inter- and intra-provider). At these sessions, teachers and educational managers examine student performances (very frequently of written language) and also examine the assessment tasks teachers have designed. The tasks contained in the CSWE are described as ‘... examples only and do not constitute a set of prescriptive, standardised assessment tasks’ (Burrows, 1993:v). However, research following the implementation of CSWE has indicated considerable variability in the tasks that teachers use to assess the same competency (see, for example Brindley 2000a, 2000b, this volume; Wigglesworth 2000a). This suggests a need for a bank of standardised tasks that can be used by teachers for assessment and as models for task design (Brindley 2000b; Wigglesworth 2000b).

**Aim and methodology of the research**

Given the need to ensure consistency in CSWE assessment, the aim of Study 2 was to examine teachers’ moderation practices with a view to determining the way in which teachers interpreted and applied the CSWE performance criteria in relation to benchmark texts. Only a small number of teachers and student performances were studied, as the emphasis was not on the outcomes of moderation, nor on quantitative indices of reliability but rather on teachers’ qualitative responses to the joint stimuli of student texts and assessment criteria. In order to do this, teachers were recorded as they undertook moderation and their commentary was examined. In this way it was hoped to establish the extent to which the raters were using the given criteria or other criteria to make their judgments. The methodology was similar to that described by Delaruelle (1997:21):

One approach to an investigation of the effects of task type on rater behaviour is through the analysis of raters’ verbal reports or protocols. During the collection of think-aloud protocols, raters verbalise their thoughts while rating a number of texts. The verbal reports are recorded, transcribed and then analysed according to a coding system.

In the present study, three student text moderation sessions and one task moderation session were recorded. In the following section, each is described and the emergent issues discussed. The outcomes of telephone interviews with the staff member who had led each moderation are then described and analysed. Finally, a number of recommendations are made with regard to moderation of classroom-based assessment within the CSWE and future directions for research are proposed.

**Moderation of student texts**

Three AM EP provider sites participated in this phase of the project. At each of the sites a group of teachers was audi-taped and videotaped during the process of moderating on a number of (written) texts. At the outset of the project it was agreed that the identity of the sites would not be revealed so that the staff would not feel constrained in their commentary. This has become increasingly important in the AM EP since the performance of all providers is increasingly monitored by DIM A. Table 9 summarises the teacher numbers, the tasks that were used and the competencies that were moderated. (The letter ‘C’ is used to indicate Competency; ‘FS’ equals Further Study and ‘CA’ equals Community Access, both study focus areas in CSWE Level III.)
same procedure for each of the three texts they moderated: after reading through the text, they marked the assessment grid for each of the performance criteria and then shared their impressions.

Student text 1, for Certificate I Competency 12, Can write a short description

This text was written in response to the prompt Describe a student in this room. The text, entitled 'Jameo', consisted of 11 sentences (some lacking capital letters at commencement). The range statement regarding length for this competency is ‘approximately 5 sentences’.

Describe a student in this room.

(Jameo)

She is Jameo. She is 34 years old. She is very beautiful. She is very pretty. She is wearing the black suite, full arm shirt and black pant. She is long black hair and wearing a black shoe on top black and white beautiful flowers. She is make her hand. She is wearing a marriage ring. She is very happy. She is very quiet. She from Jpan.

The performance criteria (pc) for this competency (NSW AMES, 1995b) are as follows:

pc 1: uses appropriate staging, eg statement of topic, description of place, person or thing
pc 2: joins simple clauses with conjunctions, eg ‘and’, ‘but’, ‘because’
pc 3: uses vocabulary appropriate to topic
pc 4: constructs simple clauses
pc 5: constructs simple noun groups and uses adjectives, eg ‘A small flat’, ‘My flat is small’
pc 6: uses ‘to be’ and ‘to have’ in the present tense
pc 7: uses personal pronouns as required.
When the teachers compared their assessments of this text, they agreed that the student had achieved all the performance criteria except for performance criteria 5 and 6. The teachers’ agreed verdict on the text was ‘Partially Achieved’.

Student text 2, for Certificate II Competency 14: Can write a short recount
This piece was written in response to the prompt: Write a one page account of the excursion to the zoo, using a model of an excursion to the Powerhouse Museum (analysed in class).

After of the teachers ask many times us about what place we want to go on friday 2 august, 1997 we decided go the taronga zoo and arrange to find our classmates and teachers at circular quy at 9:00 am. All the class were there at 9:00 am and we get into line to buy the tikets that costed $12 and inclusev trip via verry and aerial train and entrance to the zoo this was wonderfull, when we arrived at the zoo the teachers and the students was waiting for us because we leave taking photos, firstly we went to see the kangaroos, koalas and australian night life and some the most dangerous snakes, the teachers explained some of the animal types and life for us.

At 12:30 we had a muticultural picnic because every one brought different kinds of foods and drinks the lunch were nice because were natural and very original, after lunch we went to continue the excursion and went to see the seals show but I didn’t see nothing because the ground were full, when the show were finish we continued to walked around the zoo and we saw the white tiger and snow leopard that are my favorites animals. I was there for a few minutes to try to take some photos because they were sleeping and when I was out side to see the of the class I didn’t see yours for anywhere and I decided go home but when I was in the ferry station I found whit Juan antonio and he toll my that every body gone and we were the last.

All in all the excursion was very enjoyble and I knew many things about animals life and I knew new peoples and many differents foods and I’m very grateful whit the teachers because they were the people that organised the excursion tank you very much.
while another teacher observed that ‘they’ve used the strategies for structure — they’re a bit formulaic, but they’re there’.

The group verdict was that the student had only ‘partially’ achieved criterion 2, due to ‘no evidence or very tenuous (evidence)’ in some of the earlier paragraphs: a ‘whole lack of topic sentences’, ‘probably not well-substantiated’. For criterion 3, the teachers agreed that the student had demonstrated achievement. However the overuse of ‘moreover’ and ‘furthermore’ was noted. Criterion 4 was rated by the group as ‘partially achieved’, although the rating of individual criteria is designed to be pass/fail.

Criteria 5 and 6 proved decisive for this group of moderators, with comments being made on the lack of agreement between subjects and objects, the low incidence of nominalisation or formal language, the inappropriate ending, ‘very prescriptive’, and the fact that ‘the inappropriateness of the grammar and vocabulary (affects) the coherence’. The rating of this text was: ‘Partially Achieved’.

This group of moderators largely followed the procedure laid down in the guidelines for moderation (NSW AMES 1995a), with raters marking the texts separately before discussing them as a group, information which was revealed in the interview with the moderation coordinators. The group used both the performance criteria and the range statements in their examination of each text and appear to have reached decisions based on the criteria, rather than by making holistic judgments. It is interesting to note that in the last example, however, the grammar criterion would appear to have been more heavily weighted than the others, a factor also noted by Brindley (this volume).

The moderation process: Site B

Site B chose three Certificate III student texts for moderation. The texts were selected by teachers who had taught at this level recently. The moderation session proceeded along similar lines to Site A, ie the five teachers individually read through the texts, marked their assessment grids and then discussed their professional judgments.

Student text 1, for Certificate III Further Study, Competency 12: Can write a report

This piece was written in response to the following prompt: Write a factual report on your job/profession/employment field or another job/profession/employment field.
The performance criteria for this competency are:

pc 1: structures text with appropriate beginning, middle, end

pc 2: writes coherent paragraphs containing factual, closely argued information

pc 3: links ideas cohesively using, eg conjunction and reference

pc 4: uses reference to general categories as required, eg people

pc 5: uses appropriate vocabulary

pc 6: uses grammatical structures appropriately, eg simple present and other tenses as required, passive forms

When the teacher compared their ratings for this text they voiced a number of concerns. Discussion concerning criterion 1 led to agreement that the text structure was not clearly delineated: one teacher remarked on the ‘abrupt’ ending. For criterion 2, one teacher commented, ‘it’s coherent, but they’re not paragraphs’. The group agreed that the student had only ‘partially’ achieved this criterion, as ‘we’d expect them to know how to do it properly’. On the topic of ‘clearly organised information’, one teacher noted that ‘it is factual and it is clearly organised’.

Criterion 3 elicited comments regarding the use of ‘and’ in the text, with one teacher pointing out that the only two instances of ‘and’ occur at the beginning of sentences. Another teacher remarked on the shortness of a number of the sentences and the fact that most of the sentences ‘are single idea sentences’ with an uncomplicated structure. The use of ‘until’ was noted, as well as referencing through the use of a somewhat repetitive ‘it’ and a use of ‘they’. The teachers, however, agreed that the student achieved this criterion.

For criterion 4, the use of general categories such as ‘Chinese snacks’ and ‘Yum Cha’ was noted, as well as the use of reference to ‘the customer’ and the ‘cook’. This led on to a discussion as to whether there was a problem with the task in that the instruction ‘write about your profession or your job’ might tend to yield recount rather than report, as evidenced by this text. The teachers agreed that although referencing was minimal they would pass the student on this criterion.

For criterion 5, the teachers commented on such words in the text as ‘snake sales’ (for ‘snack salesman’), ‘deep fired’ (for ‘deep fried’) and agreed that there were too many errors and that they were too confusing for the student to achieve on this criterion. Finally, there was agreement on criterion 6, that the student did not demonstrate appropriate use of tenses. One teacher cited the use of future tense where present would be more appropriate and another remarked on the incorrect use of past tense in the text. The verdict on this text was ‘Partially Achieved’.

ABOUT MY JOB

Last year, I started my first job in Melbourne. I worked in Shark Fin House restaurant. It’s one of the Chinese restaurant in the city.

I am a snake sales. My job is servicing some Chinese snakes on YumCha time. It’s only on lunch time. Every weekend, from half past ten in the morning until half past three in the afternoon. It’s my working time. I go to work by tram. It takes about half hour. At first, I go to the kitchen in the restaurant, to help the cooker prepare some snakes, example, put the snakes in the plate. And then the cooker will steam or deep fired them.

After that, we had early lunch. At half past eleven, the customer will come in. They choice some snakes which they would like. And I put a mark on their bill. At half past three in the afternoon, I finish my job.
Student text 2, Certificate III, Community Access, Competency 14: Can write a short formal letter of complaint

The text (not including addresses and salutations) consisted of 87 words.

The Manager
Bargain Electrics Company
Dear Sir / Madam
I am writing to you about two weeks ago I bought the drill from Salesman of your company of home trial. I want to return the drill and I want to get refund for $15. because the drill doesn’t work and when the salesman explain that I was under no obligation to buy the drill if I didn’t like it. Also my wife points out that the same drill could be bought locally for $10 less. I hope this matter will be resolved to my satisfaction.
Yours sincerely
Student’s Name

The performance criteria for this competency are:
- pc 1: follows convention of layout for formal letter
- pc 2: stages text appropriately — beginning, middle and end
- pc 3: writes paragraphs which clearly express objective information about situations/events
- pc 4: provides information/supporting evidence; substantiates claim; requests action as required
- pc 5: uses appropriate conjunctive links, eg causal, additive, temporal, conditional as required
- pc 6: uses appropriate vocabulary to reflect topic and level of formality
- pc 7: uses grammatical structures appropriately, eg modal verbs, imperative and interrogative structures as required

The student text submitted for moderation was written in response to the following prompt: You bought an electric drill from Bargain Electrics a week ago and now wish to return it and receive a refund. Explain the reason and refer to the terms and conditions under which you bought it. Mention that you are aware that the same drill can be bought locally for $10 less.

The moderating teachers read through the text individually, completed their assessment grids and then shared their professional judgments on it. The teachers felt that the student had achieved criterion 1, allowing for the fact that there is some leeway as to how a formal letter should be addressed. One teacher felt that the student’s name above his address was inappropriate, but others overruled her on this point. The group agreed that the student achieved criterion 2 and, with regard to criterion 3, there was agreement among the teachers that it was unclear whether the drill had been purchased (as indicated in the prompt) or was on home trial (as indicated in the text). The group agreed that the student achieved criterion 4 and criterion 5, citing the use of ‘because’ and ‘also’.

For criterion 6, it was agreed that ‘there are some good phrases’, and that though these were formulaic they were appropriate. For criterion 7, one teacher noted the use of an imperative and an interrogative in wanting to award a pass, but another felt unwilling to do so. The overall verdict on this text was ‘Partial Achievement’, with a comment that ‘it leaves you a bit confused as to what happened’.

Student text 3, Certificate III Further Study, Competency 13: Can write a discussion

The text consisted of 261 words. Teacher marks on the text included ticks at the end of each paragraph and the comment ‘Excellent’ at the end.
There are several similarities between the education system in Australia and Croatia.

Firstly, both countries have three levels of education: Primary, Secondary and Tertiary. Although there are three levels of education, students do not have to obtain a Tertiary education. However, they must go to school until they are fifteen.

Secondly, children attend the school in their residential zone, usually neighbourhood, every day, Monday to Friday. This education is free, and a high school is free also. Parents must pay only for the books.

Finally, after secondary school some pupils chose to continue their education at the Tertiary level, which is for six or more years.

There are also some differences. In the first place, preschool care is optional. The parents can leave their children at home with nurse or usually with grandparents. In the Croatia children start school at the age of seven. This school is called a Secondary school. The academic year begins in September. Children also have a school brake for the summer holiday for three months, this is very different from Australia's schools.

Then, parents doesn't have to pay for the uniforms, because children doesn't wear uniforms. The first four years children have one teacher, but after that they have many teachers. One teacher for each subject.

Finally, there is no short courses last for only six months, and students must spent at five to six years at University, if they want to receive a degree in higher education.

The performance criteria for this competency are as follows:

pc 1: stages text appropriately — introduction, arguments for and against, conclusion

pc 2: writes coherent paragraphs which express arguments for and against and provides supporting evidence to substantiate claim

pc 3: uses appropriate conjunctive links, eg causal/conditional/comparative

pc 4: uses reference to signal and retrieve information as required, eg 'this means ...', 'factor ...', 'the previous ...'

pc 5: uses appropriate vocabulary with required level of abstraction

pc 6: uses grammatical structures appropriately, eg uses modal verbs, modal adverbs, clauses of condition as appropriate

The text was written in response to this prompt: Writing; Similar or different?

In about 150 words, describe in writing the similarities and differences between the Australian education system and your own. Use sentences like the following:

There are several similarities between the education system in Australia and ... Firstly, both countries ... Secondly ... Finally ... there are also some differences. In the first place: ... Then ... Finally ...

Prior to discussing their ratings for each criterion, the moderating teachers made a comment on the task, that 'it's written more as a report than a discussion' and for this reason there was 'not much for and against'. One teacher commented that the prompt, suggesting an opening using 'There are ...' was 'OK for a statement, but then you need to go on and express some argument for or against'.

It was agreed that the set task was 'quite different from the sample tasks in the Certificate' and that 'Education in My Country' would be more appropriate as a prompt for a report. It was agreed that the text should 'go back' (ie be rejected for moderation) on the grounds that the task was not appropriate for the competency and that 'the idea of opinions doesn't come through, and ways of expressing opinions aren't even suggested in the task'. Having reached this conclusion about the task, the teachers agreed that therefore it was not necessary to proceed with further moderation.

The additional point was made that the text lacked causal links and conditional links and that it needed supporting evidence, 'to make a claim about something' and that the conclusion 'needs to do some summing up of their opinion'.

At this site, the teacher-raters moderated individually before the group discussion. They used the criteria but did not seem to have referred to the range
Student text 2 consisted of a report on the ‘Northern-Eastern Tiger’ comprising 102 words. At the commencement of the moderation, the raters agreed that the text was ‘a clear pass’, remarking on ‘the clear staging’ and ‘the use of conjunctive links’. During discussion the group debated whether the wording of criterion 1 related to the use of paragraphs, or whether it related to the overall organisation of the content.

Student text 3 consisted of a report of 83 words on ‘Albury’. The group verdict was ‘a clear pass’, with teachers discussing again whether ‘staging’ related to paragraphing or to internal organisation of the text.

Student text 4 consisted of a report of 148 words entitled ‘Iraq’. The teachers’ reaction to this text was generally negative, with some feeling that it was perhaps the worst in the moderation. The view was expressed that it was largely copied from an encyclopedia, and this observation led to a discussion on the difficulties involved in monitoring Distance Learning students (ie that the amount of assistance a student is obtaining cannot be ascertained). It was pointed out, however, that teachers would be familiar with the standard of earlier work produced by the student.

One teacher raised the (rhetorical) question of ‘How far do you take into consideration the L1 of the learner? How much trouble is he going to have with particular features of the English language?’ Criterion 6 proved crucial to the group, as the text did not contain any two-clause sentences. It was agreed that the status of this text was ‘Partially Achieved’.

Student text 5 consisted of a report on ‘Parliament House’, comprising 97 words. In what appeared to be a holistic judgment, as none of the criteria was discussed individually, the teachers agreed that it was ‘a pass’ and went on to discuss whether or not the student had drawn on a brochure to compose the report. One teacher made the point that even if the student had copied some phrases he had integrated them into the report in a way that the writer of Text 4 had not done.

This moderation session revealed a number of problems both with the moderation process and with the assessment. The use of an incomplete text did not allow moderation to occur. Moreover, neither performance criteria nor range statements, which would have revealed that a number of the texts were too short, were utilised. It is also interesting to note that the session organiser allowed the moderation to proceed in this fashion without calling the criteria to the attention of the teacher-raters.
The moderation process: Site D

In this session a group of four teachers (including the session coordinator) examined reading tasks, including both texts and the questions. The group moderated two teacher-produced tasks developed for Certificate III Community Access, Competency 9: Can read an information text. The texts to be read were entitled ‘Consumer Information — Insurance Policies’ (Appendix 2) and ‘Buying A Second H and Car’.

The session commenced with a statement by the coordinator identifying the competency and the tasks to be moderated. Participants were then issued with a ‘kit’ containing:

- the requirements for the competency (from the CSWE curriculum documents) (NSW Ames 1995b)
- a “text suitability checklist” (Christie and Delaruelle 1997) (Appendix 1)
- a reading text, ‘Migraine’ from the CSWE assessment guidelines (NSW A M ES 1995b) (Appendix 2)
- two texts with questions developed by teachers.

The coordinator outlined the procedure to be followed in the session. Each text was to be read through and then checked against the ‘text suitability checklist’. If the text were deemed to be appropriate, teachers would then scrutinise the requirements for the competency contained in the CSWE curriculum documents, paying particular attention to Range Statements and the Evidence Guide. Lastly, the task would be compared with the task for this competency contained in the CSWE assessment guidelines (NSW Ames 1995b). The moderating teachers followed this procedure with both texts.

The most important issue arising from the discussion of the two texts concerned the level of difficulty. When the teachers examined, ‘Consumer Information — Insurance Policies’, there was some discussion around whether the topic would unfairly advantage some students, but consensus was reached that this would not be the case. The questions accompanying the text were then examined against the performance criteria for the competency and finally the whole task (ie text plus questions) was checked against the Assessment Guidelines task entitled ‘Migraine’. After reading though ‘Migraine’ and its accompanying questions, the teachers decided that although the text ‘Consumer Information — Insurance Policies’, was more difficult than ‘Migraine’, the questions were easier and that this struck a balance with regard to the level of difficulty. In the absence of formal trialling of tasks, this would appear to be the only method for evaluating task difficulty.

Issues raised by the moderation sessions

In the course of the student text moderation sessions at Sites A, B and C a number of issues emerged in addition to the specific issues mentioned in the description of moderation at each site. These included task validity, adherence to range statements and interpretation of the wording of some performance criteria.

Task validity

Of the ten student texts examined in the moderation sessions at Sites A, B and C, three texts demonstrated problems that could be attributable to flaws in task design. At Site A, the texts used for Certificate III Further Study, Competency 12: Can write a report, were written in response to the prompt: Write a factual report on your job/profession/employment field or another job/profession/employment field. However, the text produced was a recount of daily work activities rather than a report.

There are a number of possible explanations for the discrepancy between the task requirements and the text produced. The first is that the student misinterpreted the task. The second is that the teacher did not prepare the students for the different genres demanded by the CSWE. The third is that there may be a problem with the task which is not immediately apparent. Only trialling could assist in throwing light on this question.

The other task which caused difficulties at this site was for Certificate III Further Study, Competency 13: Can write a discussion. Here the prompt was:
Interpretation of performance criteria

During the moderation of student text sessions, the interpretation of some of the CSWE performance criteria led to considerable, and often unresolved, debate. At Site A, the questions of how accurate referencing needs to be, and the exact meaning of ‘appropriately’ were raised regarding Certificate I Competency 14:

*Can write a short recount*, pc 3: *uses simple reference appropriately*.

(It should be noted here that, in response to teacher feedback regarding its usefulness, this performance criterion was subsequently eliminated from the 1998 edition of the CSWE. See also the discussion in Brindley, this volume.)

At Site C, teachers moderating texts produced for Certificate II, Competency 15: *Can write a short report* debated whether the words ‘appropriate staging’ in Performance Criterion 1: *Uses appropriate staging* referred to paragraphs or to overall organisation of the content of the report. In the 1998 edition of the CSWE, Criterion 1 has therefore been expanded to read *Uses appropriate staging eg general classification statement*.

While the data gathered from these moderation sessions have been of use in revising the CSWE, it is significant that teachers nevertheless ‘agreed to disagree’ rather than exploring the reasons for the disagreement and seeking consensus. This would appear to put the validity of the moderation process in these sessions at risk and this issue was therefore taken up in the next stage of the research which is described below.

Interviews with moderation session organisers

In order to build up an overview of the way in which moderation sessions were arranged and conducted, the session organisers at Sites A, B, C and D were interviewed following the moderation sessions. Due to the national nature of the project and constraints of distance, a set of interviews based on a flexible questionnaire was conducted by telephone. With the participants’ consent, the interviews were recorded and transcribed.

The questionnaire was constructed to yield information on the entire process of organising and holding a moderation session, with a particular focus on the following issues:
- the role of the session organiser and how they were assigned
- the frequency and duration of sessions

Writing: Similar or different? In about 150 words, describe in writing the similarities and differences between the Australian education system and your own. Use sentences like the following: There are several similarities between the education system in Australia and ..., Firstly, both countries ... Secondly ... Finally ...
spoken performances on video. (She thought that ‘the video might have been
the NSW AMES video of benchmark performances of oral competencies’.)

The moderation session involved all twelve teachers at the school. The teachers
formed three groups and moderated the same set of texts. One group was
videotaped for the purposes of the project. The moderation session coordinator
was responsible for the selection of texts. The Certificate I and III texts were
chosen from the set of student texts moderated at the National CSWE Cross
Provider Forum (September 1997) and the Certificate II text came from her
own collection of texts. When asked if Site A ever held moderation sessions on
teacher-produced tasks, the coordinator replied that they did not hold task
moderation sessions. She went on to explain that teachers used the sample
tasks from the assessment guidelines (CSWE, NSW AMES 1995b) and that ‘we
might not use the exact model task, but we’d use the guidelines in the earlier
part of the folder’.

As moderation sessions involving more than one teacher are not the norm at
Site A, the teacher whose task it was to organise the session had no previous
experience to draw on in arranging the session. Her initial plan was for teach-
ers to moderate two texts at each Certificate Level. After discussion with the
head teacher it was agreed, however, that time constraints would allow only
one text per Certificate level to be discussed.

At the beginning of the session, teachers formed into groups based on the
Certificate level they were currently teaching. The texts and performance cri-
teria were then distributed and teachers individually assessed the texts. On
completion of this process, one group moved into another room where their
discussion of the texts was videotaped.

In response to a question as to what would happen in the event of consensus not
being reached regarding a text, the coordinator replied that in this situation a
disagreement was ‘just sort of thrashed out and the majority decision held
sway’. She made the point, also, that spoken texts had, in the past, led to greater
disagreement among teachers, with many instances of the NSW AMES verdict
on the videotaped
Student Performances for Moderation

One issue which emerged in this interview was that of time. The Site A moder-
atation coordinator made a number of references to time constraints and how
these negatively impacted on the incidence, duration and type of moderation
sessions that could be held.

Site A response

At Site A, the moderation session coordinator was simply ‘given the task’ (ie by
the head teacher) of organising it. Moderation is held once a term, usually
around week five, with a number of shorter sessions held each year. Normally,
m moderation is conducted on a one-to-one basis, with the course coordinator
allocating various competencies to different teachers on the course. The moder-
ation session was not publicised in advance.

This ‘modus operandi’ was described as follows: ‘One teacher might be given
the recount and they give the students a task and sit down with the coordinator
and work (it) out.’ The coordinator went on to comment on the time constraints
in shorter courses and concluded with the comment, ‘... even when there’s three
(teachers on a course) they don’t sit around and discuss. They each get allocated
competencies to deal with and then work it through with the coordinator.’

The moderation session held for the purpose of providing data for this project
took ‘one whole morning’, including the process of video recording, and did
not, therefore, represent the normal practice at this site.

When questioned as to whether Site A had held any moderation sessions with
other schools or organisations, the coordinator’s response was ‘not in my expe-
rience’. She went on, however, to say that some joint moderation sessions were
held with the school’s other two campuses, but that following the closure of
these campuses (approximately one year earlier), moderation had only been
conducted in the manner described above.

The focus of the moderation session at Site A was written texts. When asked
which other language domains had been focused on in the past, the coordin-
ator replied that during the previous year there had been a large moderation
session (involving one other campus) at which teachers moderated a number of

**Site B response**

In response to a question on how often moderation sessions were conducted, the Site B coordinator replied that at least two sessions were held per year ‘because we follow the quality assurance guidelines (which stipulate that) teachers have to attend at least two moderation sessions per year’. The duration of these sessions was said to be 2.5–3 hours.

At Site B, the organiser of the session explained that she had only arranged moderation sessions ‘a couple of times’ at her previous centre, and that the organisation of the session was usually carried out by the ‘curriculum person’. The procedure used was as follows. After setting the date and publicising it, the next step was to decide on the texts and/or tasks to be moderated and then to hold the session. At the end of the session, teachers filled in an evaluation sheet and then a report was made to the central curriculum support unit in that state. This report identified any ‘issues or concerns’ that might need to be further explored.

When asked whether her organisation participated in moderation sessions with other schools, the coordinator replied that it did participate, but that this was handled ‘centrally’ and she had never personally been involved in such a session.

In response to a question on the language domains usually covered in moderation sessions, writing was identified as the domain most commonly examined, with reading being described as ‘quite popular’ and listening ‘somewhat neglected’. In regard to speaking, teachers are asked to videotape their class performing specific competencies and these are then discussed. In view of the response to the survey concerning videotaping, it would be interesting to examine further how this took place and the extent to which it occurred. This issue was not pursued in this study, however.

At the beginning of the session, the ten or more teachers participating formed into groups. They then decided who would ‘report back’ to the entire group at the end of the session and proceeded to moderate the texts. When the groups had completed the moderation, they reconvened to hear the results from each group. The session coordinator listed the results and looked to see if there was ‘any commonality or differences’ amongst them.

When arranging a moderation session, the coordinator could request that teachers provide student texts and/or obtain them from the central curriculum support unit’s ‘text bank’. Deliberately, no specific type of text (eg ‘borderline’ or ‘non-achieved’) is requested, because ‘then in a way we already know what’s going to be looked for’.

In arranging the sessions, the coordinator’s first step was to set the date, which was, in her opinion, ‘a most difficult thing’ because ‘you can’t please everyone’. The advertising of the moderation session date included a reminder to teachers of their obligation to attend ‘a certain number of moderation sessions per year’. Of the 20 teachers at Site B, ten or more would attend any one moderation session.

Prior to the moderation session, texts were sometimes typed out by teachers in order to make them easier for the moderating teachers to read. All texts were then photocopied and distributed to session participants.

In reply to a question as to what happens when consensus is not reached, the coordinator replied that ‘we tend to go with the majority’. She observed that it was rare that ‘anything really poses a problem’ but that ‘we do keep an eye on people who are always out of line … being much higher or much lower than anybody else’, and the action taken was to recommend that this teacher attend more moderation sessions and to note this on the teacher’s ‘training plan’.

At Site B, moderation of teacher-produced tasks could be combined with a text moderation session or held separately. The tasks came from the assessment guidelines (CSWE, NSW AMES 1995b) or were contributed by teachers. In response to a question as to the number of tasks received, the coordinator replied that they did get ‘some’.

During the interview, the coordinator raised the question of whether all the student texts should be typed out because teachers find some students’ handwriting difficult to read. She also wondered if more tasks could be developed in the ‘neglected area’ of listening and pointed out that teachers ‘these days’ have very few hours in which to develop tasks.

**Site C response**

The coordinator at Site C explained that the role was shared between three people and ‘it doesn’t really matter who does it’. Site C had set up a series of ‘assessment folders’ into which teachers placed queries and they then got together in small groups to discuss items and ‘pass that information on to others’.
At Site C, moderation sessions were held ‘no more than once a term’, although the coordinator explained that teachers could raise problems, including requests for moderation on specific texts, at the fortnightly program meetings.

In response to a question as to whether Site C ever held moderation sessions in conjunction with other schools, the coordinator replied that they had not done so. She explained that teachers had attended moderation sessions ‘at Head Office’ and went on to say that ‘there are peculiarities about distance learning that are very special’. When asked to give examples of these differences, she described how all the students’ work was documented in their folders and that teachers went through these folders to check that various performance criteria were met and to ‘mark things off in an individual sort of way, whereas in class there’s a lot more control’.

The most common language domain moderated was writing, followed by speaking. Students who gave their permission were taped and the moderating teachers discussed the tapes. Reading was assessed ‘all the time’ through the distance learning course material, with teachers administering supplementary tasks where a course task did not cover all criteria for a competency.

Tests for moderation were selected through the following process. The site had a number of boxes in which teachers placed texts and ‘when people come across a text that might be a benchmark text, they just pop it in the box’. When there were a lot of texts in a box, ‘maybe seven or eight examples’, a moderation session was organised. In setting up a moderation session at Site C, the first step was to ascertain how many staff members would be involved: participants at Site C moderation sessions were said to number ‘about 14 or 15’. Texts in the ‘collection box’ were then counted and a decision made on which competencies would be moderated, with the group deciding whether to split into smaller groups or remain together to moderate.

Whether they stayed in one large group or broke into smaller groups, the process of moderation at Site C was the same: teachers individually assessed the texts and then discussed their ratings. In answer to a question as to what happened if teachers failed to reach agreement on the moderation of a text, the coordinator replied that ‘we agree not to reach consensus’. She added however, that ‘we’d look at the majority and … we might take into account the experience that people have in the program of working with people at a distance’.

Where there was a lack of agreement she reported that ‘we generally agreed that you look at the history of the student’. Some of the distance learning students had been in the program for one or two years and teachers ‘know them very well’. By looking back through the student’s work, the teachers could ‘easily see if something’s quite different’. This comment refers to the fact that in distance learning, unlike the classroom situation where students are writing in front of the teacher, it is more difficult to check on plagiarism, as mentioned in the section above describing the moderation sessions.

Teachers in the distance learning program developed assessment tasks to use with the Distance Learning materials, *It’s Over To You* (NCELTR 2000). These tasks were designed to ensure that performance criteria and range statements could be met by students studying by distance mode. New tasks developed by teachers were placed in the appropriate box and colleagues were invited to offer comments and suggestions. When amendments had been made to the task, it was trialled. Then, when ‘everyone is happy’, the task was placed in Site C’s assessment task bank.

At various points in the interview the coordinator made the point that distance learning is very different from the classroom style of teaching and learning and that this had had a considerable impact on how moderation sessions were conducted.

**Site D response**

At Site D, the moderation session involved an examination of two teacher-produced reading tasks for Certificate III, Community Access, Competency 9: Can read an information text. Since the focus of the session was on teacher-produced tasks rather than student texts, the interview with the Site D coordinator took a somewhat different form from the interviews with the coordinators at Sites A, B and C.

At Site D, moderation of student texts took place at least twice a year, with each teacher attending two (whole day) sessions per year. Moderation sessions lasted three hours each, with morning sessions focusing on Certificate I and the competencies in Certificate II which were designated to belong in Phase 1, that is the ‘easier’ competencies, and afternoon sessions covering Certificate II, Phase 2 and Certificate III. The coordinator reported that moderation sessions focusing on teacher-produced tasks were held ‘sometimes twice a year’. However, she went on to say that in the previous year no such sessions had been held.
What is a moderation session?

Of major importance is the need for a common understanding of what actually constitutes a moderation session. In discussions with teachers, it appeared that the concept of informal, collegial discussion generally involving one or two other teachers was regarded as ‘moderation’. While this consultation process is extremely valuable and widely used, as discussed in Part One, it should not be seen as equal to or able to supplant a formally organised moderation session involving a number of texts or tasks and a group of teachers.

How many texts should be moderated?

The question of the number of texts to be moderated at a session is also of interest. In the moderation sessions that were videotaped for this study, two of the three venues involved conducted the moderation session using one student text per competency. However, the practice of moderating one text only does not allow teachers to see a full range of performances. They are thus unable to develop an overall view of the competency and the standards that are involved in the assessment of a number of performance criteria.

What happens when consensus is not reached?

As discussed above, in a situation where teachers do not agree on their verdict of achieved/not achieved for a performance criterion or a whole text, the tendency is for them to decide to ‘agree to disagree’. While it is acknowledged that in competency-based assessment, 100 per cent agreement is not always achievable it would, however, seem beneficial for teachers to have recourse to an advisory panel for assistance when controversy arises. (In this regard, it is common practice in high profile language proficiency tests to seek a third rating when there is significant disagreement between two raters.) Teachers would then be able to develop an enhanced understanding of the standards involved in the achievement or non-achievement of a competency. In addition, where the texts are of interest in highlighting interpretations of performance criteria or range statements, they could be used in other moderation forums.

Conclusion and recommendations

The results of this study suggest that there are a number of aspects of CSWE assessment and moderation that are in need of further attention. First, the apparent lack of systematic double rating merits closer scrutiny. It should be acknowledged that since the AMEP was put to tender, it has been essential for

The recorded session was arranged in response to teachers who felt that there were not enough tasks available for this particular competency. These teachers then designed some tasks and the session was arranged to establish whether the tasks produced were ‘at a satisfactory level and whether they met the performance criteria’. Prior to the session, the coordinator telephoned centres to locate teachers with an interest in task design who would be available to participate in the moderation session and four teachers participated.

In Site B’s state, the preferred method regarding teacher-produced assessment tasks was for tasks to be sent in to the central curriculum support unit. They were then examined and a moderation session was set up for teachers to evaluate the tasks according to the performance criteria for the competency.

Prior to the session, the coordinator photocopied the following resources:

a) her own checklists of ‘things that they need to look at’ regarding text suitability

b) the checklist from the set of resources developed to assist teachers in designing CSWE assessment tasks (Christie and Delaruelle 1997)

c) the competency, Certificate III, Community Access, Competency 9: Can read an information text competency sheet from the Certificates in Spoken and Written English

d) a task for this competency from the CSWE assessment guidelines. (Appendix 2).

When they examined the teacher-produced tasks and compared them with the checklists given them by the coordinator, the teachers agreed that the questions did in fact ‘globally’ address the performance criteria.

The coordinator made the point that the curriculum support unit had a ‘good bank’ of assessment tasks and that teachers ‘feel more comfortable to work with assessment tasks which have actually been moderated’. While this process was not mandatory, teachers were encouraged to participate in it. If a task was deemed satisfactory and the teacher/s producing it were willing to share it, the task was then sent to all centres in the state.

Organising moderation sessions: Emergent issues

A number of issues emerged from the analysis of the four moderation sessions. These issues can be examined as a series of questions.
educational providers to minimise the cost of delivering the Program in order to compete effectively for funding. In such a climate, it is difficult to recommend that second raters be made routinely available for all assessment. Nevertheless, the variability in teachers’ interpretations of the CSWE performance criteria noted in this study highlights the subjectivity of the rating process and points to the desirability of involving more than one rater. The problems inherent in the ‘live’ rating of oral interaction would be alleviated by regular audiotaping and videotaping of spoken language performances. This would provide a concrete basis for collegial discussion and help teachers to develop a common understanding of the performance criteria that characterise different Certificate levels.

The failure of some benchmark texts to elicit the required text type and the differing interpretations of performance criteria highlight the need for a set of assessment tasks which meet the CSWE requirements for the competency in question and which can be used as exemplars in moderation sessions (see also Brindley 2000b, this volume). In view of the differing practices revealed by this study, there would also appear to be a need to provide professional development in the conduct of moderation exercises, possibly in tandem with a new edition of the Christie and Delaruelle (1997) moderation kit. These initiatives could be undertaken in conjunction with the next edition of the CSWE. Given that moderation sessions often resulted in participants ‘agreeing to disagree’, there is also a need for an ‘advisory panel’ to which moderators could appeal for a verdict on ‘problematic’ texts.

In summing up — this study raises a number of larger issues and unresolved dilemmas in relation to the validity and reliability of curriculum-related assessment. Clearly, it would be hard to argue that the assessment practices described here would meet the technical requirements of a high profile proficiency test. However, is it reasonable to expect that a classroom-based assessment should be subjected to the same degree of scrutiny as a well-funded high stakes international test? On the other hand, what are the implications of not applying these criteria to classroom-based assessment? As Clapham (2000) argues, the results of classroom assessments may have a significant impact on individual learners, even if these assessments appear to be ‘low stakes’. If this is so, then there is clearly a case for ensuring that CSWE assessment is carefully developed and adequately resourced. At the same time, however, practicality is an extremely important consideration: the development and maintenance of high quality assessment systems come at a cost and this cost needs to be balanced against the benefits to learners. Finally, in considering the direction of assessment policy, the question of impact is crucial. The types of learners studying within the CSWE framework include Aboriginal people living in remote communities, survivors of torture and trauma, refugees and members of various disadvantaged groups. For such learners, a rigorous, centralised testing regime would clearly be inappropriate and could not be recommended, despite the difficulties involved in classroom-based assessment.

The dilemma thus remains of how to put in place an assessment system which directly reflects the curriculum, is fair to learners and provides high quality information, but which is neither totally standardised and test-driven nor affordable with the resources available.

The answer might partly be found in Radnor and Shaw’s reconciliation model (1995:139ff), which uses a combination of ‘insider’ (teacher) and ‘outsider’ (external assessor) perspectives to arrive at a consensus ‘about how work is to be valued and criteria interpreted’ (p142). This approach would help teachers to identify and maintain a set of agreed standards in regard to achievement at different Certificate levels. To obviate the potential negative effects on learners of formal testing, various forms of ‘non-intrusive’ assessment can be used to collect ongoing evidence of learners’ progress and achievement as part of normal classroom activities. Such procedures include portfolios, structured observations, journals, self and peer assessment and project work (Brown and Hudson 1998). Coupled with the types of moderation procedures described above, the use of these less formal procedures would go some way towards providing an ongoing record of competency achievement that could be used for reporting purposes while at the same time enhancing the quality of learning.

References


Acknowledgments

The project team would like to thank the following people for their assistance and participation in the project:

Joyce Murray (NSW AMES)
Wendy Mrowka (NSW AMES)
Christine Howell (AMES, Victoria)
Eva Grieve (AMES, Victoria)
M arion M artin (ELLS)
M aree Fieldes (UNSWIL)
M eg Qu inn (TAFE Queens land Language and Literacy Services)
Peter M cGuire (NTU)
Christine Bowden (WA AMES)
Liz Stone (NSW AMES)
Lili Lippa (AMES, Victoria)
Anne Quested (NSW AMES)
Linda Ross (NSW AMES)
Sandra Economou (WA AMES)
Cathy Domahidy (WA AM ES)
Antoinette M ulders (NSW AM ES)
Raymond Leung (N CELTR)

In addition, we would like to thank all the people who participated in the project in their state AM ESs.


2

Investigating rater consistency in competency-based language assessment

Geoff Brindley

Introduction

One of the most significant challenges in implementing any assessment system that requires subjective judgments of human performance is ensuring that these judgments are consistent. Lack of consistency leading to measurement error may result in erroneous inferences concerning learners' abilities and culminate in unfair decisions that may affect individuals' lives (Hughes 1989). For this reason, many educational institutions and testing agencies make a considerable investment in rater training and reaccreditation in order to allow raters to familiarise themselves with the rating criteria and to practise applying them to samples of performance.

In the national Australian Adult Migrant English Program (AMEP), assessment of learners' achievement of language competencies within the Certificates in Spoken and Written English (CSWE) is carried out by teachers on the basis of learners' performance in oral interaction, reading and writing, using specific linguistic criteria (see Brindley 2000a, 2000b, 2000c; Burrows, this volume). In view of the critical role of teacher assessments in certification and the need for consistency identified above, it is important to investigate not only the level of agreement between teacher-raters on the quality of learner performance samples but also to explore the extent to which the performance criteria are interpreted and applied in the same way (for further discussion see Brindley 2000d).

The study reported in this chapter aims to throw some light on these questions. It investigated the consistency with which experienced CSWE assessors were able to classify samples of spoken and written production corresponding to three oral interaction competencies and six writing competencies. Rating
patterns of a national group of 12 experienced CSWE assessors are analysed using quantitative techniques, supplemented where relevant with reference to the raters’ comments on the rating process and the texts provided. Possible reasons for inconsistencies in rating are discussed and recommendations are made for improving CSWE assessment practices.

**Method**

AMEP provider organisations across Australia were requested to submit samples of oral and written production for one oral competency and two writing competencies from each of Certificates I, II and III and, where possible, to identify those samples which they considered to be ‘benchmarks’, that is, those which illustrated a minimum standard of achievement of the competency in question. The samples were coded and numbered according to their origin. A total of 22 spoken videotaped spoken performances and 72 written texts were received. The competencies represented by the samples are listed below.

**Oral interaction competencies**

Samples of texts representing one competency at each of the three Certificate levels were used as follows.

**CSWE I**
- Competency 4 Can provide personally relevant information using spoken language (six samples from two locations)

**CSWE II**
- Competency 5 Can give spoken instructions (eight samples, two locations)

**CSWE III**
- Competency 6 Can negotiate a complex/problematic spoken exchange (eight samples, four locations)

**Writing competencies**

The 12 texts representing samples of performance in each competency were as follows.

**CSWE I**
- Competency 12 Can write a short description (seven locations, nine benchmark texts)
- Competency 13 Can write a short recount (seven locations, six benchmark texts)

**CSWE II**
- Competency 14 Can write a short recount (six locations, seven benchmark texts)
- Competency 16 Can write a short opinion text (four locations, six benchmark texts)

**CSWE III (Further Study)**
- Competency 13 Can write a discussion (nine locations, nine benchmark texts)

**CSWE III (Community Access)**
- Competency 14 Can write a short formal letter of complaint (seven locations, nine benchmark texts)

**The rating process**

A group of 12 experienced CSWE assessors was assembled, representing nine different AMEP provider organisations in seven Australian states and territories. The assessors were asked to rate all of the oral and written performance samples according to each of the performance criteria using a scale of 0 to 2, where 0 = not achieved, 1 = achieved and 2 = high achieved. A randomised rating design was used to minimise order effects.

**Data analysis**

One of the main aims of the project was to identify the extent to which ‘benchmark’ texts deemed by organisations to represent examples of minimal competency achievement were classified in a similar way by experienced assessors working in different locations. The percentage of raters (PR) who certified the competency in question as having been achieved was therefore first calculated. In this way it was possible to ascertain the extent to which the national group broadly agreed with the original text classification. In accordance with the CSWE rating scheme, a spoken or written text was deemed to be achieved if it received a score of one or more on each of the performance criteria. The award of one or more ratings of zero on any of the criteria automatically meant that the competency was classified as not achieved.

Scores on each of the performance criteria for each text were summed to yield a score for each text from each assessor. A one-way analysis of variance with repeated measures was then carried out on the total scores. This yielded a mean score, standard deviation and standard error for each text and enabled any significant differences between texts and between assessors to be identified.
The analysis also provided a generalisability coefficient (G-coefficient), analogous to the classical internal consistency reliability estimate. This statistic gives an indication of the extent to which raters agree on the ranking of candidates.

In the case of the writing competencies, the scores awarded by different assessors to individual performance criteria within each competency were also examined in order to investigate the extent of agreement between raters and to ascertain whether the criteria differed notably in difficulty. This analysis was carried out by calculating each rater's total score and the mean scores awarded for each performance criterion within each writing competency. The range of scores awarded to each performance criterion was also tabulated.

Many of the assessors wrote free-form comments on the rating sheets concerning the quality of the texts provided and their application of the performance criteria. The same issues were also the subject of a general discussion following the rating exercise. These comments were examined and drawn on where necessary in order to clarify reasons for assessors' scoring decisions.

Results

The results of the analysis of ratings for each text representing the oral and written competencies are shown below. For both the spoken and written competencies, the mean ratings for each text are first presented, along with the standard deviation (SD), standard error (SE) and percentage of raters who agreed with the classification of the text as submitted (PR). Where texts were identified by organisations as ‘benchmarks’, they are marked with an asterisk.

The second table which accompanies each of the writing competencies presents the mean scores and range of scores on the performance criteria within each writing task.

The commentary which follows will focus on the extent to which raters were able to agree on the classification of texts, looking particularly at the degree of agreement on texts which had already been classified by assessors as ‘benchmark’ level prior to the rating session described here. In cases where inconsistent rating patterns are identified, a range of possible explanations will be explored.

Oral competencies

Table 10: Certificate 1 Competency 4: Can provide personally relevant information using spoken language

<table>
<thead>
<tr>
<th>Text</th>
<th>Mean</th>
<th>SD</th>
<th>SE</th>
<th>PR</th>
</tr>
</thead>
<tbody>
<tr>
<td>B2</td>
<td>2.000</td>
<td>0</td>
<td>0</td>
<td>1.000</td>
</tr>
<tr>
<td>A17A</td>
<td>1.580</td>
<td>.90</td>
<td>.26</td>
<td>.667</td>
</tr>
<tr>
<td>A17B</td>
<td>2.000</td>
<td>0</td>
<td>0</td>
<td>1.000</td>
</tr>
<tr>
<td>A17C</td>
<td>2.580</td>
<td>.52</td>
<td>.149</td>
<td>1.000</td>
</tr>
<tr>
<td>A17D</td>
<td>1.833</td>
<td>.718</td>
<td>.207</td>
<td>.833</td>
</tr>
<tr>
<td>A17E</td>
<td>2.080</td>
<td>.515</td>
<td>.515</td>
<td>.917</td>
</tr>
<tr>
<td>AVE</td>
<td>2.011</td>
<td>.351</td>
<td>.101</td>
<td>.905</td>
</tr>
</tbody>
</table>

Overall agreement between assessors on the classification of the texts for this competency was very high, with perfect or better than 90% agreement on five of the six texts. However, the ANOVA revealed significant differences between texts (F = 5.892, p = .0001) and between assessors (F = 3.306, p = .001), and the generalisability coefficient was at the borderline of acceptability at .697. This suggests that although the judges were able to broadly agree on the status of the texts, they were less unanimous in their rankings.

Texts B2, A17B and A17C elicited complete agreement amongst assessors. All were classified as benchmark samples of performance or better. Of these Text A17C was rated the highest at 2.58, with seven assessors giving the first performance criterion: answers simple questions eg. requests for personal details a rating of ‘high achieved’. Post-hoc Scheffé tests revealed a significant difference between the ratings for this text and those awarded to text A17A which was rated by four assessors as ‘not achieved’.

Text B2 elicited a number of comments from assessors. In relation to the performance criterion: asks for repetition or clarification as required, five assessors commented that the student shown on the sample videotape did not have the opportunity to ask questions, although all of them rated this criterion as having been fulfilled.
Table 11: Certificate 2 Competency 5: Can give spoken instructions

<table>
<thead>
<tr>
<th>Text</th>
<th>Mean</th>
<th>SD</th>
<th>SE</th>
<th>PR</th>
</tr>
</thead>
<tbody>
<tr>
<td>B2</td>
<td>3.66</td>
<td>.492</td>
<td>.142</td>
<td>.667</td>
</tr>
<tr>
<td>E1A</td>
<td>1.917</td>
<td>1.128</td>
<td>.326</td>
<td>0.000</td>
</tr>
<tr>
<td>E1B</td>
<td>4.167</td>
<td>.389</td>
<td>.112</td>
<td>1.000</td>
</tr>
<tr>
<td>E1C</td>
<td>3.75</td>
<td>.866</td>
<td>.250</td>
<td>.917</td>
</tr>
<tr>
<td>E1D</td>
<td>4.667</td>
<td>.651</td>
<td>.188</td>
<td>1.000</td>
</tr>
<tr>
<td>E1E</td>
<td>2.08</td>
<td>1.165</td>
<td>.336</td>
<td>.167</td>
</tr>
<tr>
<td>E1F</td>
<td>3.25</td>
<td>1.29</td>
<td>.372</td>
<td>.500</td>
</tr>
<tr>
<td>E1G</td>
<td>4.083</td>
<td>.90</td>
<td>.260</td>
<td>1.000</td>
</tr>
<tr>
<td>AVE</td>
<td>3.458</td>
<td>.54</td>
<td>.156</td>
<td>.657</td>
</tr>
</tbody>
</table>

Assessors were in complete agreement on the classification of four of the eight texts for this competency. However, there was a wide range of ratings on the other texts. No assessor considered text E1A to meet the benchmark standards and only two rated text E1E as an example of achievement. Four assessors commented on the poor sound quality of the videotape provided with the latter text and/or on the poor pronunciation of the student concerned. Opinion was evenly divided on text E1F, with six assessors classifying the competency as achieved and the other six as not achieved.

There was no significant difference between assessors but post-hoc tests revealed significant differences between a number of the texts ($F = 18.179$, $p = .001$). Text E1A was rated significantly lower than all other texts with the exception of E1E which only two assessors classified as meeting the standard. The latter text was rated significantly lower than texts B2, E1B, E1C and E1D. There was also a significant difference between text E1F and text E1D. The generalisability coefficient was very low at .32, reflecting a substantial difference in the ranking of the texts.

Table 12: Certificate 3 Competency 6: Can negotiate a complex/problematic spoken exchange

<table>
<thead>
<tr>
<th>Text</th>
<th>Mean</th>
<th>SD</th>
<th>SE</th>
<th>PR</th>
</tr>
</thead>
<tbody>
<tr>
<td>B2</td>
<td>5.167</td>
<td>.577</td>
<td>.167</td>
<td>1.000</td>
</tr>
<tr>
<td>B1A</td>
<td>5.333</td>
<td>.778</td>
<td>.225</td>
<td>.917</td>
</tr>
<tr>
<td>B1B</td>
<td>4.417</td>
<td>.900</td>
<td>.260</td>
<td>.416</td>
</tr>
<tr>
<td>H1A</td>
<td>3.917</td>
<td>1.084</td>
<td>.313</td>
<td>.416</td>
</tr>
<tr>
<td>H1B</td>
<td>3.75</td>
<td>1.215</td>
<td>.351</td>
<td>.416</td>
</tr>
<tr>
<td>H1C</td>
<td>4.083</td>
<td>1.379</td>
<td>.398</td>
<td>.500</td>
</tr>
<tr>
<td>H1D</td>
<td>3.833</td>
<td>.835</td>
<td>.241</td>
<td>.167</td>
</tr>
<tr>
<td>H1E</td>
<td>4.333</td>
<td>.985</td>
<td>.284</td>
<td>.583</td>
</tr>
<tr>
<td>AVE</td>
<td>4.354</td>
<td>.473</td>
<td>.137</td>
<td>.552</td>
</tr>
</tbody>
</table>

There was a good deal of disagreement between assessors on the texts selected as examples of this competency. Only one text, B2, was unanimously classified as representing competency achievement, while 11 of the 12 assessors agreed on text B1A. However, on five of the eight texts which were rated, opinion was almost evenly divided as to whether the texts met the minimal standard. Significant differences were found between texts ($F = 4.903$, $p = .0001$), and assessors ($F = 1.969$, $p = .0398$). The generalisability coefficient was low at .492.

Two assessors commented that students appeared to be reading the cue cards in the tasks for competencies H1B and H1D.
Writing competencies

Table 13: CSWE Certificate 1 Writing Competency 12: Can write a short description

<table>
<thead>
<tr>
<th>Text</th>
<th>Mean</th>
<th>SD</th>
<th>SE</th>
<th>PR</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1A</td>
<td>9.083</td>
<td>2.644</td>
<td>.763</td>
<td>.083</td>
</tr>
<tr>
<td>A13*</td>
<td>7.25</td>
<td>2.006</td>
<td>.579</td>
<td>.167</td>
</tr>
<tr>
<td>A5*</td>
<td>5.333</td>
<td>.888</td>
<td>.256</td>
<td>.000</td>
</tr>
<tr>
<td>B1A</td>
<td>6.417</td>
<td>1.505</td>
<td>.434</td>
<td>.000</td>
</tr>
<tr>
<td>D1A</td>
<td>7.167</td>
<td>2.167</td>
<td>.626</td>
<td>.167</td>
</tr>
<tr>
<td>D1B</td>
<td>7.000</td>
<td>2.089</td>
<td>.603</td>
<td>.167</td>
</tr>
<tr>
<td>C1B*</td>
<td>5.333</td>
<td>1.723</td>
<td>.497</td>
<td>.000</td>
</tr>
<tr>
<td>B2*</td>
<td>5.417</td>
<td>1.165</td>
<td>.336</td>
<td>.000</td>
</tr>
<tr>
<td>A18*</td>
<td>10.000</td>
<td>2.663</td>
<td>.769</td>
<td>.833</td>
</tr>
<tr>
<td>C1C*</td>
<td>2.833</td>
<td>3.129</td>
<td>.903</td>
<td>.000</td>
</tr>
<tr>
<td>C1D*</td>
<td>6.75</td>
<td>2.667</td>
<td>.770</td>
<td>.083</td>
</tr>
<tr>
<td>B1B*</td>
<td>5.583</td>
<td>1.311</td>
<td>.379</td>
<td>.000</td>
</tr>
<tr>
<td>AVE</td>
<td>6.514</td>
<td>1.025</td>
<td>1.025</td>
<td>.125</td>
</tr>
</tbody>
</table>

It can be seen that of the eight texts submitted as benchmarks for this competency, only one text, A18, was classified by a majority of assessors as representing minimal attainment. Of the remaining benchmark texts, five (A5, C1B, B2, C1C and B1B) were unanimously classified as examples of non-achievement, although not consistently on the basis of the same performance criteria. The other two texts, A13 and C1D, were deemed to be achieved only by one and two raters out of 12.

Significant differences were found between texts (F = 11.373, p = .0001) and assessors (F = 2.17, p = .0191). The generalisability coefficient was fairly low at .539.

Text C1C, with a mean rating of 2.833 (significantly lower than the overall mean rating of 6.514) appeared to be particularly problematic and elicited a range of comments. Four assessors expressed the view that this text was of ‘the wrong genre’ and therefore did not allow students to demonstrate the required features of writing. One assessor, however, rated the text as ‘high achieved’ on six of the eight performance criteria. Texts C1A, A13, D1A, D1B, A18 and C1D also elicited a good deal of disagreement. Text A18, with a mean rating of 10, received a number of ratings of 2 from many assessors and could therefore be seen as an example of high achievement, although two assessors considered that it did not meet all of the performance criteria. Although text C1A, with a high mean rating of 9.083, received ratings of 2 from nine assessors on one or more performance criteria, only one assessor considered that it met all the criteria.

Table 14: CSWE Certificate 1 Writing Competency 12: Can write a short description

<table>
<thead>
<tr>
<th>Performance criteria</th>
<th>Mean</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Staging</td>
<td>8.833 (0.736)</td>
<td>4-13</td>
</tr>
<tr>
<td>Clauses with ‘and’</td>
<td>6.583 (0.549)</td>
<td>5-9</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>13.667 (1.139)</td>
<td>12-16</td>
</tr>
<tr>
<td>Simple clauses</td>
<td>12.75 (1.063)</td>
<td>7-17</td>
</tr>
<tr>
<td>Simple noun groups</td>
<td>10.00 (0.833)</td>
<td>3-17</td>
</tr>
<tr>
<td>‘Be’ and ‘have’</td>
<td>8.160 (0.674)</td>
<td>4-15</td>
</tr>
<tr>
<td>Personal pronouns</td>
<td>11.500 (0.959)</td>
<td>4-15</td>
</tr>
<tr>
<td>Prepositional phrases</td>
<td>6.750 (0.563)</td>
<td>1-12</td>
</tr>
</tbody>
</table>

The second performance criterion joins simple clauses with ‘and’ was rated quite low, with a mean rating of .549. There was also some inconsistency in the ratings on this criterion, with less than two-thirds agreement amongst assessors on six of the 12 texts. These inconsistencies may have been due to a liberal interpretation on the part of some assessors who noted that students were able to join sentences by other means and therefore accepted connectives other than ‘and’ as meeting the criterion, whereas others applied the criterion as stated.

The fifth criterion: constructs simple noun groups incorporating an article, an adjective and a noun was, similarly, rated very low, with a mean rating of .563, suggesting that this feature may be harder than others for students to produce. It also attracted a very wide range of ratings (from 3 to 17). One assessor remarked that ‘one noun group should be adequate’ and two others commented that the article was not always necessary in a description.

The sixth performance criterion in this competency: uses ‘to be’ and ‘to have’ in the present tense, and the final criterion: uses some prepositional phrases to indicate location (which received a very wide range of ratings) received low mean ratings, with means (for single ratings) of .674 and .563 respectively. The former elicited a number of comments from four assessors, several of whom felt that some of the texts provided met the requirements for an adequate
As with Competency 12, the performance criteria relating to clauses with ‘and’ and simple noun groups were rated lowest, with means of 0.875 and 0.563 respectively. The former received a very wide range of ratings, suggesting problems in interpretation and/or application of this criterion.

Table 15: Certificate I Competency 13: Can write a short recount

<table>
<thead>
<tr>
<th>Text</th>
<th>Mean</th>
<th>SD</th>
<th>SE</th>
<th>PR</th>
</tr>
</thead>
<tbody>
<tr>
<td>A18A</td>
<td>7.75</td>
<td>1.712</td>
<td>.494</td>
<td>0.083</td>
</tr>
<tr>
<td>A18B</td>
<td>7.917</td>
<td>1.73</td>
<td>.499</td>
<td>0.000</td>
</tr>
<tr>
<td>A19A</td>
<td>10.333</td>
<td>2.902</td>
<td>.838</td>
<td>1.000</td>
</tr>
<tr>
<td>A13*</td>
<td>7.75</td>
<td>1.815</td>
<td>.524</td>
<td>0.167</td>
</tr>
<tr>
<td>A18C*</td>
<td>8.083</td>
<td>1.832</td>
<td>.529</td>
<td>0.500</td>
</tr>
<tr>
<td>C2</td>
<td>6.083</td>
<td>.996</td>
<td>.288</td>
<td>0.000</td>
</tr>
<tr>
<td>A20A*</td>
<td>10.25</td>
<td>2.563</td>
<td>.740</td>
<td>0.917</td>
</tr>
<tr>
<td>A19B*</td>
<td>11.00</td>
<td>2.045</td>
<td>.590</td>
<td>0.833</td>
</tr>
<tr>
<td>A7*</td>
<td>5.25*</td>
<td>1.138</td>
<td>.329</td>
<td>0.000</td>
</tr>
<tr>
<td>A19C</td>
<td>11.417*</td>
<td>1.975</td>
<td>.570</td>
<td>0.583</td>
</tr>
<tr>
<td>A20B</td>
<td>8.333</td>
<td>1.557</td>
<td>.449</td>
<td>0.833</td>
</tr>
<tr>
<td>A5</td>
<td>6.667</td>
<td>1.073</td>
<td>.31</td>
<td>0.000</td>
</tr>
<tr>
<td>AVE</td>
<td>8.403</td>
<td>1.136</td>
<td>.328</td>
<td>0.285</td>
</tr>
</tbody>
</table>

Only one text for this competency, A19A, was deemed by all assessors to be of benchmark standard, although it had not been submitted as such. Of the five texts designated as benchmarks, two — A20A and A19B — were classified as achieved by 10 or 11 of the 12 raters. Of the others, opinion was evenly divided on whether text A18C met the minimum standard, only two assessors considered text A13 to be an example of achievement and not one assessor rated text A7 as achieved.

Significant overall differences were found between texts (F = 19.59, p = .0001) and assessors (F = 3.033, p = .001). The generalisability coefficient was moderate at .67. The mean score for text A7 was significantly lower than the overall mean rating and text A19C significantly higher. Texts A20A, A19B and A19C received ‘high achieved’ ratings on a number of criteria from most assessors. However, despite receiving the highest overall score, text A19C was classified as an example of achievement by only 7 of the 12 assessors.

Three assessors noted that Texts A19B and A19C were ‘almost identical’. One commented that ‘these two recounts must have been written after a class session where the teacher went over the whole excursion on the board’.

Table 16: C SW E I W riting Competency 13: Can write a short recount

<table>
<thead>
<tr>
<th>Performance criteria</th>
<th>Mean</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Staging</td>
<td>12.417 (1.035)</td>
<td>7–17</td>
</tr>
<tr>
<td>Clauses with ‘and’</td>
<td>10.500 (0.875)</td>
<td>4–17</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>13.667 (1.194)</td>
<td>12–20</td>
</tr>
<tr>
<td>Simple clauses</td>
<td>12.750 (1.104)</td>
<td>7–19</td>
</tr>
<tr>
<td>Simple noun groups</td>
<td>10.000 (0.563)</td>
<td>5–10</td>
</tr>
<tr>
<td>Action verbs</td>
<td>15.917 (1.326)</td>
<td>12–21</td>
</tr>
<tr>
<td>Personal pronouns</td>
<td>12.667 (1.056)</td>
<td>12–17</td>
</tr>
<tr>
<td>Time or location</td>
<td>14.917 (1.243)</td>
<td>10–18</td>
</tr>
</tbody>
</table>

Table 17: Certificate II Competency 14: Can write a short recount

<table>
<thead>
<tr>
<th>Ratings</th>
<th>Text</th>
<th>Mean</th>
<th>SD</th>
<th>SE</th>
<th>PR</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVE</td>
<td>5.597</td>
<td>.673</td>
<td>.194</td>
<td>0.535</td>
<td></td>
</tr>
<tr>
<td>A18B</td>
<td>5.583</td>
<td>.793</td>
<td>.229</td>
<td>0.750</td>
<td></td>
</tr>
<tr>
<td>A15*</td>
<td>4.750</td>
<td>.965</td>
<td>.279</td>
<td>0.250</td>
<td></td>
</tr>
<tr>
<td>A20A*</td>
<td>4.000</td>
<td>1.044</td>
<td>.302</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>G1A</td>
<td>4.833</td>
<td>1.267</td>
<td>.366</td>
<td>0.500</td>
<td></td>
</tr>
<tr>
<td>G1B</td>
<td>5.833</td>
<td>1.115</td>
<td>.322</td>
<td>0.583</td>
<td></td>
</tr>
<tr>
<td>A18B</td>
<td>7.667</td>
<td>1.875</td>
<td>.541</td>
<td>0.917</td>
<td></td>
</tr>
<tr>
<td>D1A*</td>
<td>5.583</td>
<td>1.564</td>
<td>.452</td>
<td>0.583</td>
<td></td>
</tr>
<tr>
<td>A18C</td>
<td>7.250</td>
<td>.965</td>
<td>.279</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>D1B*</td>
<td>4.833</td>
<td>.937</td>
<td>.271</td>
<td>0.250</td>
<td></td>
</tr>
<tr>
<td>A20B*</td>
<td>5.833</td>
<td>.835</td>
<td>.241</td>
<td>0.583</td>
<td></td>
</tr>
<tr>
<td>A18D*</td>
<td>5.083</td>
<td>1.379</td>
<td>.398</td>
<td>0.500</td>
<td></td>
</tr>
</tbody>
</table>

As with Competency 12, the performance criteria relating to clauses with ‘and’ and simple noun groups were rated lowest, with means of .875 and .563 respectively. The former received a very wide range of ratings, suggesting problems in interpretation and/or application of this criterion.
No one of the seven texts submitted as benchmarks yielded complete agreement. The majority of assessors (7 out of 12) considered that texts D1A and A20B met the benchmark standard and six assessors classified texts C1 and D1B as examples of achievement. Only three assessors considered texts A15 and D1B to meet all criteria for achievement. Text A20, provided as an example of minimal achievement, was not classified as completely achieved by any of the assessors.

Significant differences were found between texts ($F = 12.14, p = .0001$) and assessors ($F = 3.109, p = .0009$). The generalisability coefficient was moderate at .678. The mean score for text A18B, classified as achieved by 11 of the 12 assessors, was significantly higher than the mean overall rating and this text can be seen as an example of high achievement on this competency. Texts G1A, G1B, D1A and A18D provoked the most disagreement.

Six of the 12 assessors noted that Text A18A was too short to fulfill the range statements. Nine made the same comment in relation to text A20 and eight in relation to Text A15.

Of the five texts designated as benchmarks, only text A19F was deemed by all assessors to demonstrate achievement of all criteria. This text, with a mean rating of 7.333, was rated significantly higher than the overall mean rating of 4.674. Only one of the other four benchmark texts, B2, was deemed to be achieved by more than half the assessors. A sizeable majority of assessors did not agree with the benchmark classification for texts A19D, A19E and A19G. Significant differences were found between texts ($F = 17.69, p = .0001$) and assessors ($F = 2.304, p = .0125$). The generalisability coefficient was low at .566.

Nine of the 12 assessors commented that text A18A was too short to meet the requirements of the range statement. Nine made the same comment in relation to text A20 and eight in relation to Text A15.

The last criterion: constructs some sentences containing two clauses received the lowest mean rating, suggesting it was the hardest to meet. Three assessors commented that this criterion was too demanding and that it should read ‘one or more’ clauses.

One assessor wrote that the performance criteria for this competency at Certificate Two level were ‘not as prescriptive as CSWE 13’ and that the competency ‘doesn’t seem to expect much more of a higher standard’. One commented that the performance criteria were ‘simpler than pcs for Stage 1’.

<table>
<thead>
<tr>
<th>Table 18: CSWE II Writing Competency 14: Can write a short recount</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Performance criteria</strong></td>
</tr>
<tr>
<td>Staging</td>
</tr>
<tr>
<td>Conjunctive links</td>
</tr>
<tr>
<td>Simple reference</td>
</tr>
<tr>
<td>Vocabulary</td>
</tr>
<tr>
<td>Past tense and markers</td>
</tr>
<tr>
<td>Two-clause sentences</td>
</tr>
</tbody>
</table>

The last criterion: constructs some sentences containing two clauses received the lowest mean rating, suggesting it was the hardest to meet. Three assessors commented that this criterion was too demanding and that it should read ‘one or more’ clauses.

Of the five texts designated as benchmarks, only text A19F was deemed by all assessors to demonstrate achievement of all criteria. This text, with a mean rating of 7.333, was rated significantly higher than the overall mean rating of 4.674. Only one of the other four benchmark texts, B2, was deemed to be achieved by more than half the assessors. A sizeable majority of assessors did not agree with the benchmark classification for texts A19D, A19E and A19G.

Significant differences were found between texts ($F = 17.69, p = .0001$) and assessors ($F = 2.304, p = .0125$). The generalisability coefficient was low at .566.

Nine of the 12 assessors commented that text A19A was too short to meet the requirements of the range statement and two assessors considered the assessment task to be inappropriate. Six assessors made the same comment about text A19B. Five assessors suggested that tasks associated with some or all of texts A19A, A19B, A19C, A19D and A19E did not produce the required type of writing since they elicited a report or an interpretation of data rather than an opinion.

<table>
<thead>
<tr>
<th>Table 19: CSWE II Competency 16: Can write a short opinion text</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Text</strong></td>
</tr>
<tr>
<td>A19A</td>
</tr>
<tr>
<td>A19B</td>
</tr>
<tr>
<td>A19C</td>
</tr>
<tr>
<td>A19D*</td>
</tr>
<tr>
<td>A19E*</td>
</tr>
<tr>
<td>B2*</td>
</tr>
<tr>
<td>A19F*</td>
</tr>
<tr>
<td>A19G*</td>
</tr>
<tr>
<td>A19H</td>
</tr>
<tr>
<td>C2</td>
</tr>
<tr>
<td>A12</td>
</tr>
<tr>
<td>A19J</td>
</tr>
<tr>
<td>AVE</td>
</tr>
</tbody>
</table>
Of the nine texts which had been designated as benchmarks by organisations submitting them, only one, A7A, was deemed by a majority of assessors (in this case 8 out of 11) to meet all of the performance criteria. Of the other benchmark texts, text B1A received an overall rating of ‘achieved’ from 5 out of 12 assessors, text A7B was rated as achieved by 5 out of 11 assessors, text H1 by 4 out of 11 and texts A8, B1B, A13 were classified as achieved by 3 assessors out of 11. Two of the benchmark texts, G1B and E1, received an overall classification of ‘not achieved’ from all assessors.

There were significant differences between texts ($F = 13.275$, $p = .0001$) and assessors ($F = 4.143$, $p = .0001$), although the generalisability coefficient was .759, suggesting reasonable agreement in ranking of the texts. Text G1A received very low ratings, with a mean rating of .583 compared to an overall mean of 4.361. In this context, eight assessors commented that this text was too short to fulfil the range statements. Text A13 was also judged to be too short by four assessors.

### Table 22: CSWE III Writing Competency 13: Can write a discussion

<table>
<thead>
<tr>
<th>Performance criteria</th>
<th>Mean</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Staging</td>
<td>8.364 (0.760)</td>
<td>5-13</td>
</tr>
<tr>
<td>Coherent paragraphs</td>
<td>7.909 (0.659)</td>
<td>3-12</td>
</tr>
<tr>
<td>Conjunctive links</td>
<td>9.545 (0.795)</td>
<td>4-15</td>
</tr>
<tr>
<td>Reference</td>
<td>7.456 (0.621)</td>
<td>1-12</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>10.636 (0.886)</td>
<td>7-16</td>
</tr>
<tr>
<td>Grammatical structures</td>
<td>7.417 (0.777)</td>
<td>3-14</td>
</tr>
</tbody>
</table>

The most difficult category in this competency was reference (uses reference to signal and retrieve information as required, e.g., ‘this means …’, ‘this factor …’, ‘the previous …’). Along with uses appropriate conjunctive links, this performance criterion elicited a very wide range of ratings. Vocabulary was the easiest feature in this competency.
Table 23: CSWE III Competency 14: Can write a short formal letter of complaint

<table>
<thead>
<tr>
<th>Text</th>
<th>Mean</th>
<th>SD</th>
<th>SE</th>
<th>PR</th>
</tr>
</thead>
<tbody>
<tr>
<td>A2*</td>
<td>7.667</td>
<td>1.775</td>
<td>0.512</td>
<td>0.833</td>
</tr>
<tr>
<td>A17A</td>
<td>7.000</td>
<td>2.374</td>
<td>0.685</td>
<td>0.750</td>
</tr>
<tr>
<td>A15*</td>
<td>7.500</td>
<td>2.812</td>
<td>0.812</td>
<td>0.417</td>
</tr>
<tr>
<td>A8A*</td>
<td>4.500</td>
<td>1.679</td>
<td>0.485</td>
<td>0.167</td>
</tr>
<tr>
<td>A8B*</td>
<td>4.75</td>
<td>1.913</td>
<td>0.552</td>
<td>0.167</td>
</tr>
<tr>
<td>A8C*</td>
<td>3.883</td>
<td>1.030</td>
<td>0.297</td>
<td>0.000</td>
</tr>
<tr>
<td>A8D*</td>
<td>3.917</td>
<td>1.730</td>
<td>0.499</td>
<td>0.000</td>
</tr>
<tr>
<td>A10*</td>
<td>2.917</td>
<td>1.240</td>
<td>0.358</td>
<td>0.000</td>
</tr>
<tr>
<td>A17B</td>
<td>2.917</td>
<td>2.021</td>
<td>0.583</td>
<td>0.083</td>
</tr>
<tr>
<td>G1A*</td>
<td>5.167</td>
<td>1.642</td>
<td>0.474</td>
<td>0.167</td>
</tr>
<tr>
<td>G1B*</td>
<td>7.083</td>
<td>2.906</td>
<td>0.839</td>
<td>0.583</td>
</tr>
<tr>
<td>G1C</td>
<td>5.417</td>
<td>1.048</td>
<td>0.417</td>
<td>0.250</td>
</tr>
<tr>
<td>AVE</td>
<td>5.222</td>
<td>1.048</td>
<td>0.302</td>
<td>0.285</td>
</tr>
</tbody>
</table>

Of the nine texts submitted as benchmarks for Competency 14, only two — A2 and G1B — were classified by a majority of assessors (10 and 7 out of 12 respectively) as meeting all of the criteria for competency achievement. The remainder of the benchmark texts were rated as non-achieved by a significant majority of assessors. Texts A8C, A8D and A10 were not considered by any assessor to meet the criteria for achievement of this competency.

Significant differences were found between texts (F = 11.976, p = .0001) and assessors (F = 2.718, p = .0032). The generalisability coefficient was .632.

There were substantial differences across raters in the scores awarded to texts G1B, A15, A17A and A17B, suggesting not only disagreement between assessors on whether a particular performance criterion had been achieved or not, but also differing interpretations of what constituted high achievement, with three raters awarding ratings of 2 on this competency with much greater frequency than others.

Table 24: CSWE III Witing Cmpetency 14: Can write a short formal letter of complaint

<table>
<thead>
<tr>
<th>Performance criteria</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Layout</td>
<td>11.417 (0.951)</td>
<td>10–15</td>
</tr>
<tr>
<td>Staging</td>
<td>11.583 (0.965)</td>
<td>9–15</td>
</tr>
<tr>
<td>Paragraphs</td>
<td>7.750 (0.646)</td>
<td>5–12</td>
</tr>
<tr>
<td>Supporting evidence</td>
<td>10.000 (0.833)</td>
<td>7–16</td>
</tr>
<tr>
<td>Conjunctive links</td>
<td>8.25 (0.688)</td>
<td>2–15</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>7.16 (0.598)</td>
<td>2–11</td>
</tr>
<tr>
<td>Grammatical structures</td>
<td>6.25 (0.520)</td>
<td>2–12</td>
</tr>
</tbody>
</table>

Uses grammatical structures appropriately was the most harshly marked performance criterion, eliciting, along with uses appropriate conjunctive links and uses appropriate vocabulary to reflect topic and level of formality, a wide range of scores. The amount of variation in scoring these criteria suggests that the notion of ‘appropriacy’ as far as these features are concerned is not shared across raters, even though specific examples are given in the assessment grids. Can use appropriate staging and layout were the easiest criteria for learners.

Summary

The analysis of the ratings of the benchmark texts in this study was based on raw scores and thus was unable to take account of raters’ systematic tendencies to rate consistently severely or leniently across all of the tasks. In this regard, it would have been preferable to have used many-facet Rasch analysis to derive adjusted estimates based on the entire set of ratings (Linacre 1989), as in a previous study of CSWE writing competencies (Brindley 2000d). However the numbers of ratings were insufficient here to justify the use of the Rasch model.

Although the analysis is necessarily somewhat less fine-grained than would have been possible using an item response model, a number of broad conclusions can nevertheless be drawn from the analysis of the ratings. In the first place, it appears that the national group of assessors did not consistently agree that those writing texts which had been previously classified by AMEP providers as benchmarks were in fact examples of minimal competency achievement. Second, the assessors showed a good deal of variability both in their overall classification of the sample texts and in their interpretation and application of individual performance criteria. Third, the analysis suggested that both the texts and the individual performance criteria were of differential difficulty, both within and across the competencies.
Investigating rater consistency in competency-based language assessment

Believed that the learner had produced an acceptable example of the text type required, whereas others strictly adhered to the criterion as specified in the competency statement. In the example cited above, it would appear that some raters were willing to rate the performance criterion as achieved if they observed that the learner had used other means of connecting sentences. The same phenomenon was noted in a number of other competencies (cf. the discussion above of the descriptive text required by CSWE 1 Competency 12 where a number of raters considered it was possible to write a successful descriptive text without the use of an article or the verb ‘to have’). This raises the question of the extent to which it is possible to classify a given set of linguistic features as being obligatory (as opposed to optional) for the successful performance of any task (see Quinn 1993; Brindley 1994, 1998 for further discussion).

A second source of variation in the ratings is the texts themselves. Here a number of the tasks were identified by raters as not eliciting examples of the genre required and therefore classified as ‘non-achieved’. Given that some of these texts were intended to represent ‘benchmark’ samples of minimal achievement, this suggests that not all AMEP providers have a shared understanding of the particular features which define a given piece of writing.

Improving consistency: Future directions

Identifying suitable texts and tasks

Given what might appear to be subtle differences between some genres (note, for example, raters’ comments on differences between an opinion text and a description in CSWE II Competency 16), there is clearly a need to build up a bank of texts which are agreed examples of the type of writing required and a set of tasks which have been found to effectively elicit the performance criteria. Clear guidelines identifying the key features of the text that is required, accompanied by examples of acceptable and non-acceptable tasks, would help to reduce uncertainty amongst teachers and make the assessment process fairer for learners (see Claire, this volume, for further discussion and recommendations).

Addressing performance criteria

Although it is very difficult to develop performance criteria that are subject to unambiguous interpretation, it seems clear from this study that raters find some of the CSWE criteria particularly difficult to interpret and apply. The same picture also emerges from Smith’s (2000) study of CSWE writing assessments.
Conclusion

This study involved an examination of raters’ consistency in classifying oral and written texts as ‘achieved’ or ‘not achieved’ using the CSWE performance criteria. On the basis of a quantitative analysis of rating patterns supplemented by the raters’ own comments, some inconsistencies in the interpretation and application of the performance criteria have been identified.

In the light of these findings, a number of avenues have been suggested which could be pursued in order to improve the quality of the assessment tasks and rating procedures that are used (see also Brindley 2000d; Wigglesworth 2000). Further detailed recommendations for improving consistency in data collection and task moderation are given by Claire (this volume). In the future, ongoing monitoring and evaluation of the Program’s assessment practices will be needed to ensure that assessment is able to provide the valid and reliable information on learner outcomes that is required by all AMEP’s stakeholders. Ideally the progressive monitoring and assessment would be accompanied by systematically targeted professional development and further research into the way in which assessment tasks are constructed and designed.

References


Appendix 1

Text suitability checklist
(Christie and Delaruelle 1997:44)
Is a text suitable?

When trying to decide if a text is suitable for use in an assessment task, you should consider the following questions.

1. Will the topic of this text discriminate against or offend certain individuals or groups within the assessment population because of its cultural bias, gender bias, religious bias?

2. Will the topic of the text arouse a strong emotional response in some candidates which may adversely affect their performance eg AIDS, organ donation etc?

3. Is the topic so technical that many candidates will feel overwhelmed by the technical language in the text eg. quantum physics?

4. Is the topic in such a specialised field that it may advantage candidates coming from that particular field who may have a more sophisticated understanding of the topic than that presented in the text, eg. latest developments in computer technology/the greenhouse effect?

5. Is the information in the text so well known that candidates might have a high degree of success on the questions just by using their general knowledge of the topic, eg. reading instructions for an automatic train ticket machine? Students should not be able to answer the questions without reading the text or by relying on prior knowledge.

6. Is the text selected for the assessment task the correct text type?
   For example, the instructions on a Disprin packet would be more appropriate as an information text rather than a procedural text. NB: Narratives need to have the appropriate staging: orientation ▲ complication ▲ resolution.

7. Does the text contain a number of specialised, technical or generally difficult lexical items that may be off-putting to students?
   If this is the case but you like the text and want to use it, substitute simpler words.

8. Are the layout and graphics of a written text easy to follow?
   If not, the text may need to be retyped or reformatted.

9. Is the text an appropriate length?
   Often authentic materials will need to be shortened or simplified in order to make them suitable for a particular CSWE level.

10. If a text has been written by a teacher, is it too simplistic and contrived?

11. Is the text well written?
   It is often hard to write effective items if the text is poorly written. In particular, newspaper articles need to be carefully screened because the amount of editing that occurs means that the texts are often disjointed and reference features within the text are not obvious.

12. Does the text contain a lot of colloquial language which may be beyond the grasp of the students?
   This may be the case with some newspaper articles, yet these texts can lose their distinctive character if the colloquial language is modified.

13. Is the text interesting? Would you be interested in reading it yourself? If not, forget it. We don’t want students to fall asleep during the assessment task.
Appendix 2

CSWE III reading assessment tasks used in moderation sessions
Consumer Information

Insurance Policies

Most people have insurance policies for cars, their house and their furniture. Many people have insurance policies against damage to or loss of valuable belongings, such as jewellery. But insurance policies can be very complicated and confusing documents, many of the complaints made to consumer protection agencies concern insurance policies.

Many of these problems wouldn't happen if consumers knew more about the type of policy they were buying, about the way to make an insurance claim or about how to make a complaint if they felt unhappy about the whole process.

To help consumers buy insurance policies that suit their needs - and to help them understand the policies when they're buying them - the Federal Bureau of Consumer Affairs has developed this information resource which explains the process and ways through some of the possible pitfalls.

What to look for when buying insurance

Ask yourself: “Why do I need insurance?” When you've got the answer clear in your mind, you should shop around, read all the documents carefully, and only then buy the policy that suits your needs.

Read more than just the insurance company brochure to find out what a policy covers you for and what your rights and responsibilities under it will be. You must carefully read the policy document - that's the contract with the insurance company - before you actually sign it so that you'll be clear about what you'll pay and what you'll get for your money.

What to look for in the policy document

The policy document should be in plain everyday English with no hard-to-understand legal terms and no fine print. You should be able to understand every sentence in it.

It should contain clear statements about the circumstances in which the insurance company won't pay your claim. For example, some home contents insurance policies cover loss of jewellery worn outside the home; others don't.

It should contain clear statements which tell you exactly how to make a claim and include details: about any time limits for making claims. It should also tell you how much it will cost you to make claims and about the maximum claim that you can make.

It should be clear, precise and helpful, not only about your rights under your policy, but also about your responsibilities, which are just as important.

Some Do's and Don’ts when buying Insurance

- **Don't** accept verbal assurances.
- **Do** notify the insurance company in writing about changes to your policy circumstances and **do** get clear written confirmation from the company about these changes.
- **Do** keep all correspondence with your policy documents.
- **Do** declare all items of value to be insured, even if the policy document doesn't require you to do so. You should supply enough detail to identify items or to prove purchase.
- **Do** keep copies of this information with your policy documents.
- **Do** get the insurance company to give you clarification in writing if you aren't sure about any of the conditions of an insurance policy.

What to do if you need to make a complaint

You should first contact the insurance company itself.

If you can't sort your problem out with the company, you should then contact **Insurance Enquiry and Complaints Ltd Australia (IEC)** on **1300 363 683**.
Consumer Information — Insurance Policies

Student Name _______________________________________________

1. Who produced this information for consumers?
____________________________________________________________________________________________________________________________________________________________

2. What are the two main reasons this information was produced?
____________________________________________________________________________________________________________________________________________________________

3. Under what heading is information on who to contact if you have problems?
____________________________________________________________________________________________________________________________________________________________

4. If you are not sure about any of the conditions of the insurance policy, what should you do?
____________________________________________________________________________________________________________________________________________________________

5. Now that you have read the information, list three things you would look for in an insurance policy document.
____________________________________________________________________________________________________________________________________________________________

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.
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
PART 2
3
Searching for washback: The impact of assessment in the Certificate in Spoken and Written English
Catherine Burrows

1 Introduction
The study reported in this section examines the Adult Migrant English Program's implementation of a new assessment system that was introduced as part of a competency-based curriculum, the Certificates in Spoken and Written English (CSWE). The adoption of the Certificates provided an ideal opportunity to examine the impact of a new assessment system on the curriculum. However, this study differed somewhat in its focus from most washback studies which conventionally investigate the effects of large-scale standardised ‘curriculum-free’ proficiency tests, in that it set out to examine the impact of an assessment system that was classroom-based and closely connected to the curriculum.

The structure of Part Two of this volume is as follows. First, it describes the context in which the study was undertaken and situates the present study in relation to other washback studies. It then describes the impact on teachers of the implementation of CSWE assessment through an analysis of survey, interview and observation data. Finally, it examines the conceptualisation of washback current at the time the study was undertaken and proposes a new model for washback.

The context of the study
The climate of change
This study took place during an era of change both in Australia and internationally, when ‘... teaching (and, for that matter, education itself)’ was ‘in a state of profound flux and change’, with teachers describing a ‘... sense of rapid, overwhelming and (sometimes) meaningless change ...’ (Australian Teaching Council 1995:6).
The impact of assessment in the Certificate in Spoken and Written English

However, the national and international climate of educational and economical accountability is impacting on AMES in a number of different ways... AMES (NSW) is facing demands to provide... detail on just what learning outcomes are being achieved... (Manidis 1992:10)

The CSWE was part of NSW AMES’ response to the Australian Commonwealth Government’s Training Reform Agenda (Joyce 1993:7; Hood 1995), an agenda which saw “life-long learning” as both the solution to long-term unemployment and as the key to workplace reform and restructuring (Free 1994a; Free 1994b). This emphasis on workplace reform and skills acquisition dominated educational thinking in Australia during the 1990s.

Competency-based training (CBT) and criterion-referenced assessment were key elements of the educational reform that was sweeping Australia. These were being widely adopted both in industry and in initial teacher education (Hallinan and Danaher 1994). CBT arose from a desire amongst educationalists to measure performance in terms of the meeting of specific, task-based objectives and to report outcomes that described student competence (Glaser and Reynolds 1964; Lindvall 1964; Bormuth 1970; Mager 1975). It is expressed in terms of the following formula by Romiszowski (1981:182):

### Competency based education = mastery learning + modular individualized instruction.

Melton (1994:285–6) described the advent of CBT in Great Britain where, by the mid-1980s, vocational education had already found its way into schools-based education, under an initiative designed to equip 14-18-year-olds more effectively for employment (Dale 1985). In Australia, the National Training Board established a set of competency-based assessment principles (National Training Board 1992:71) and described competency-based assessment as ‘... a viable business and personal development tool ...’ (National Training Board 1993:24).

Competency standards in Australia were seen to have been generated by government policy and thus part of an explicitly political agenda that linked education narrowly to economic productivity (Chappell 1996:59). In this regard, CBT was seen by some commentators as a tool of economic rationalism (Toms 1995). Chappell (1996:62) criticised the narrowness of the CBT model which he saw as being...
... based on the view that standardised training outcomes can be achieved by all learners if a thorough analysis of the behaviours demonstrated by any competent performer is undertaken and then transposed into a set of standardised learning sequences.

At the time when the CSWE was being designed and implemented, CBT was seen to hold great promise by some in Australia (Mawer 1992; Docking 1993; Guthrie 1993; Rumsey 1993; Burrows 1994). Over a considerable period, however, a wide range of concerns had been raised about the potential for CBT to be associated with behaviourist views of learning (Mawer 1975:23). These included:

- the danger of confusing ‘... clarity with a high degree of specificity’ (Tyler 1964:78), that is, learning is not necessarily enhanced merely by outlining the tasks to be achieved through a highly specific set of performance criteria;
- the potential atomism of the approach (Hager 1996:74; Adult Literacy Information Office 1996);
- the ‘... extent to which assessment shapes the teaching and learning process under a competency-based training system ...’ (Robinson 1993:24) that is, the potential negative washback effects of competency-based assessment.

Others were not convinced of the benefits of CBT (Kenway, in Reid 1992; Walker 1993; Toms 1995) or did not believe that the adaptation of CBT and its assessment to language learning was appropriate (Quinn 1993; Grove 1997). Quirk (1994:39) canvassed many of the issues involved in the interface between language and literacy teaching and CBT, and questioned the extent to which competency-based assessment could meet the needs of different audiences:

Is it possible for competency-based assessment to reconcile conflicting requirements ... namely:

- the expectation that assessment in adult education should be informal and non-threatening, administered in the context of a negotiated curriculum, facilitate learning and provide constructive feedback to learners,
- and the expectation that assessment will be useful for accountability, external reporting and certification? (Quirk 1994:39)

Brindley (1994:41) described CBT as dominating vocational education in Australia, the UK and New Zealand. He pointed to a lack of evidence for the construct validity of the CSWE competency statements (p 46), and identified some of the problems involved in using a binary achieved/not achieved assessment scheme. Brindley also raised a number of issues relating to the generalisability of the CSWE, including the difficulties of making generalisations concerning student ability on the basis of a single performance and the lack of comparability of assessments administered by different teachers (ibid pp 49–50). Many of these concerns echo the issues raised in relation to performance assessment in general education (see, for example Messick 1992).

Quinn saw great similarity between CBT in America and in Australia (1993:55). He also claimed that CBT was not suitable for refugees, on the basis that it was behaviourist and atomistic (ibid p 58). Others, however, did not agree (Williams 1994; Hood 1995). Williams (1994:10) argued that

The CSWE identifies relevant criteria and contexts for the competencies it identifies, and in this respect provides qualifications to broadly defined competencies. These are very different from the functionally-based competencies used in the US ... which don't look familiar to Australian teachers.

Assessment in the AMEP

Before the introduction of the CSWE, the only formal assessment tool used in NSW AMES was the NSW AMES Oral Proficiency Rating Scale (NSW AMES 1979), a scale used in conjunction with an oral interview that was used to place students into classes. In other states, the Australian Second Language Proficiency Rating (ASLPR) scale (Ingram and Wylie 1984) was used for this purpose. In 1989, the results of a survey of AMEP teachers' assessment practices by Brindley (1989) were published, giving a clear picture of the practices that were in use before the introduction of CSWE. Brindley’s findings suggested that, aside from oral proficiency rating, almost no formal assessment of students was taking place within the AMEP (Brindley 1989:31). Not all AMEP learners were satisfied with this situation, however. Alcorso and Kalantzis (1985:148) reported that 80 per cent of the learners they surveyed ‘... expressed a strong preference for the greater use of more formal means of assessment such as “tests or formal examinations” ...’ Students surveyed also perceived the potential for assessment to be used as a means of accessing some form of accreditation or certification (ibid pp 135–6). These results were supported by studies carried out by Kessler (1984a; 1984b) and Burton (1991).
Teachers, too, felt the need for some means of assessing and reporting their students’ progress. In a survey of AMEP teachers’ views on curriculum and assessment undertaken in 1987, Nunan (1987:31) reported a collective view that a ‘... general communicative test which tested qualitative rather than quantitative outcomes could have a beneficial effect on the curriculum’ although teachers were concerned about the form such assessment might take. On the other hand, they felt that ‘... an objective standard against which learner achievement would be assessed ... could lead to a “return to the numbers game”’ (ibid).

Assessment in the CSWE

When the CSWE was introduced into AMEP classrooms, it was accompanied by a set of Assessment Guidelines (Burrows 1993a). The assessment approach reflected in these guidelines was based on the following underlying principles:

- Assessment of achievement should be competency-based, and criterion-referenced achievement.
- Students should be explicitly informed when they were being assessed.
- Reliability was to be achieved by ‘... close attention to ... the performance criteria and range of variables’. To this end, studies of the reliability of the assessment of CSWE began during the first year of implementation (see, for example, Burrows 1993b) and have been regularly undertaken since that time (Claire, this volume; Brindley, this volume).
- Assessment was to be carried out at the end of a sequence of work, when the teacher felt confident that the majority of students being assessed would be successful.
- A student was to be given as many opportunities as practicable to achieve the competencies required, although only one attempt was required if the student was successful.
- While the credential was formal, the method of assessment did not need to be. Observation and continuous monitoring could be used to gather evidence of achievement.
- The student was required to achieve all the performance criteria within a competency to achieve that competency and each performance criterion was equally weighted (Burrows 1993a:vii–ix).

The Assessment Guidelines provided ‘... guidance on assessment principles, procedures, tasks and record keeping in relation to the assessment of competencies across all stages of the curriculum framework’ (Burrows 1993a:v). They included sample assessment tasks accompanied by assessment grids allowing teachers to record students’ progress against the performance criteria. The use of these grids and tasks was not mandatory, although the assessment of the students against the competencies was.

Teachers were instructed to ‘present items similar to’ the sample assessments tasks which had been developed to elicit the features of the text type required by the competency in question. They were advised that ‘students should be very familiar with the competency and have done a number of similar tasks before being formally assessed’ (Burrows 1993a:viii).

Unlike many standardised tests which are separated from the learning process and aim to gather information on proficiency that is acquired independently of a particular curriculum, the assessment of CSWE was intended to represent ‘... a model of assessment which is to be formative, diagnostic, summative, evaluative and criterion-referenced’ (Troman 1989:288). Rather than being a tool for summative measurement of outcomes, the assessment tasks were designed to be closely connected to the learning process and to provide information that could be used by teachers and learners for purposes of motivation and improvement. This relationship between assessment and instruction in the CSWE is described by Brindley (1998:52) thus:

... since there is a direct link between attainment targets, course objectives and learning activities, assessment is integrated closely with instruction: what is taught is directly related to what is assessed and (in theory at least) what is assessed is, in turn, linked to the outcomes that are reported.

The fact that the competencies represented both the curriculum objectives and the assessment outcomes meant that the further this research went, the more the ‘connectedness’ of the assessment was revealed, a view which was found to be shared by the majority of the participants in the research.

To examine the washback effect in the context of classroom- and competency-based assessment, it is first necessary to define washback. The following chapter will describe how the definition of washback changed for language testers and curriculum writers, particularly during the 1990s, as empirical research methodologies were applied and previous assumptions were undermined.
The impact of assessment in the Certificate in Spoken and Written English

2 The washback hypothesis

In recent years researchers in the field of language assessment have turned their attention to the effects of assessment on the teaching and learning process, known as ‘washback’. In their seminal article ‘Does washback exist?’, Alderson and Wall (1993) describe and discuss the claims for washback made by educational researchers. Most commonly, washback was defined as ‘... the direct or indirect effect of examinations on teaching methods’ (Prodromou 1995:13) and ‘... the impact tests have on teaching and learning ...’ (Shohamy 1993b:4). Other definitions of washback outlined an interrelationship with student motivation (Paris et al 1991; Fransson 1984); a connection between the stakes of a test and its impact (Gipps 1994); and a distinction between testing and assessment (Gipps 1994). The latter emphasis is uncommon in the washback literature but most pertinent to this study.

Foucault’s view was that the examination is ‘... the most efficient tool through which society imposes discipline ...’ (1979, in Shohamy 1993a:49). This is supported by Hargreaves who described the ‘... imposition of standardized testing to control what teachers teach ...’ as a device ‘... for changing teachers ...’ underpinned by ‘... the presumption that teachers have somehow fallen short, and that intervention by others is needed to get them up to scratch’ (Hargreaves 1994:11–12).

Bailey described ‘a basic model of washback’ in which a test impacted on participants (students, teachers, materials writers, curriculum designers and researchers) and products (learning, teaching, new materials, new curricula and research results) with each element of the latter group impacting on one another (Bailey 1996:264). She developed this concept by dividing washback into ‘learner washback’ and ‘program washback’ (Bailey 1997:24), and explored each in depth. She also described other participants in washback, including test developers, teacher educators, curriculum planners, teacher advisers, administrators ‘... and even parents ...’ (ibid: 50–1).

Within the literature, there appears to be a clear divide between those who believe washback to be negative, most commonly characterised by the claim that teachers were forced by the introduction of the test to ‘teach to the test’; and those who believe that the introduction of a new test could have beneficial results, a view most commonly heard in the context of ‘authentic’ or ‘performance’ tests. Claims that washback had a negative influence on students, teaching and educational systems were extensive (Vernon 1956; Marris 1961; Valette 1967; Wilkinson 1968; Forbes 1973; Madsen 1976; Frederiksen 1984; Hughes 1989; Haladyna et al 1991:2; Smith 1991; Brown 1992; Shohamy 1993a; Fullan 1993:51; Gipps 1994:37, 52–3). Most claims for negative washback concerned a narrowing of the curriculum, student anxiety and a restriction of time available for teaching.


Before 1993, claims for both negative and beneficial washback were almost always predicated on an assumption of the existence of washback, rather than on the basis of research demonstrating its existence. Such unsubstantiated claims made it possible for researchers and educationalists to advance opinion as fact, so heralding the use of a new test or demanding that a test be modified or abandoned.

The washback hypothesis

The work of Alderson and Wall in the Sri Lankan Impact Study (Wall and Alderson 1993) and their paper, ‘Does washback exist?’ (Alderson and Wall 1993), had a major influence on the concept of washback in language testing.
and particularly on the methodology used to undertake research in this area in the mid to late 1990s. They posed the fundamental question about the lack of empirical evidence for the existence of washback, for ‘washback validity’ (Morrow 1986:6) or for ‘systemic validity’ (Frederiksen and Collins 1989:27), a view supported by Mesick (1996).

In order to demonstrate the multifaceted nature of concepts of washback, Alderson and Wall listed 15 possible phrasings of the ‘Washback Hypothesis’, from the most general: ‘A test will influence teaching’, to the most specific: ‘Tests will have washback effects for some learners and some teachers, but not for others’ (1993:120–1). However, general definitions of washback such as these were not sufficiently precise to answer Alderson and Wall’s challenge to define washback exactly and provide empirical evidence for its existence (Alderson and Wall 1993:128).

During the 1990s a number of studies into washback using empirical methods was undertaken. Following the publication of Alderson and Wall (1993) and Wall and Alderson (1993), the value of observation as a research tool in washback studies quickly became obvious. In this context, Bailey (1996:273) comments:

The most complete designs for researching washback include both observing classes and asking the participants about their views and experiences (either through interviews or written questionnaires) to determine whether teaching and learning are ‘evidentially linked to the introduction and use of the test’. (Mesick 1996:242)

This view was supported by other washback studies where empirical methods, including observation, had been used successfully to reveal new facets of washback (for example Khaniya 1990; Watanabe 1992; Alderson and Wall 1992; Blewchamp 1994; Watanabe 1996; Alderson and Hamp-Lyons 1996).

Because claims of washback ... have been based almost wholly on opinions rather than on empirical data ... we saw the observation of actual TOEFL preparation classes as the most important component of our data. (Alderson and Hamp-Lyons 1996:287)

Watanabe (1996:332) stated that perhaps the most important implication of his research was that ‘... we must not take for granted unproven assertions about washback’. By gathering empirical evidence, Watanabe (1992) was able to cast doubt on commonly held beliefs about the Japanese examination system. In a similar vein, Shohamy used classroom observation to examined the effects of the introduction of three tests and uncovered new complexities in washback:

All three tests had some type of impact. The impact is complex, occurring in a number of directions, and is strongly dependent on the nature and purpose of the test; it also changes over time. (Shohamy 1993b:14)

Alderson and Wall (1993:41) contended that ‘... the Washback Hypothesis ... is in need of considerable refinement’ and it was this refinement which they had attempted to undertake in their major study of the impact of the Sri Lankan O-Level examination. From 1989 to November 1991, trained observers reported on the effects on Sri Lankan classrooms of the new O-Level English examination and a new communicative textbook series which were introduced in order to force change on English language teaching practices. Despite teachers in their study consistently stating that they felt that their teaching had been affected by the examination, the researchers found that ... observers’ accounts ... suggest that there is little independent examination impact on content ... and that there is no examination impact on methodology. The impact on content grows as the exam approaches ... (However) It is not clear that they are comfortable with or in many cases even aware of what ... the new O-Level exam is trying to test. (Alderson and Wall 1993:181–2)

Of all the studies of washback in the literature, the Sri Lankan O-Level project was most like that undertaken by the researcher in this study of the AMEP. These studies had in common the complexity of the naturalistic situations being researched, unlike other studies which involved an experimental design. Each also had a complicating factor, in the case of the Sri Lankan study, the introduction of the textbooks (English for Me and English Every Day); and, in the case of the AMEP, the implementation of the curriculum, which the assessment was designed to support. Both, therefore, were studies of assessment changes which accompanied classroom changes. Both studies took place in a climate of political change, although the governmental and policy changes experienced in Australia cannot be compared with the serious political difficulties encountered in Sri Lanka.
Testing versus assessment

Unlike most washback research, the emphasis in the present study was on classroom-based, teacher-constructed, ‘connected’ assessment and not on a standardised, national or international proficiency test. In this context, Gipps (1994) had considered the question of a difference between testing and assessment and, like Shohamy et al (1996), had commented on the link between washback and the stakes of the tests involved. The stakes of the assessment of CSWE increased over time as it was adopted more and more widely and eventually became the mandatory basis for reporting student outcomes to DIMA. The assessment of CSWE, however, did not carry the high stakes of tests such as TOEFL. The students’ view of the stakes is unknown, since it was decided to restrict the study to teachers, because of the emerging importance of the teacher in washback and because the researcher had limited access to the students who had studied in the Program prior to the implementation.

Washback and educational change

The similarity between washback and other types of educational change suggested the need for a re-evaluation of washback in the context of classroom-based assessment. Wall (1996) was one of the few to examine the washback effect on individuals and to explore similarities between washback and other types of change. The fact that studies of curriculum innovation and washback have grown from different areas of educational research, which, in turn, had different purposes, may explain why the possibility of a similarity between the two areas had not previously been more fully explored. In the field of language teaching and testing, tests of language proficiency have traditionally been designed for specific purposes such as university admission, with concomitant potential changes in classroom practice (at least in the minds of some writers); while curricular innovation has been described as ‘... a managed process of development whose principal products are teaching (and/or testing) materials, methodological skills, and pedagogical values that are perceived as new by potential adopters’ (Markee 1997:46).

Because of the overlapping of curriculum and assessment concerns, an issue for this study was whether in fact the implementation of the assessment of CSWE was actually a question of curriculum innovation or educational change and not of washback. When one reads the descriptions of washback such as those...
The impact of assessment in the Certificate in Spoken and Written English


The rationale for this strategy is that the flaws of one method are often the strengths of another, and by combining methods, observers can achieve the best of each, while overcoming their unique deficiencies.1

Ex post facto research

In conducting any investigation of the effects of change on beliefs, attitudes or behaviour, it is clearly preferable to be able to track implementation of the change from the point at which it is introduced. Unfortunately, however, in the complex and often messy world of educational practice, this is not always possible and researchers into curriculum innovation and change frequently have to gather data retrospectively (see Beretta and Davies 1992 for an example of a language program evaluation using reconstructed historical data).

The present study also faced this problem. Since the curriculum and its assessment had already been implemented before the study began, it was necessary to undertake ex post facto research (Merriam 1988:7). This involved the collection of information on past events which could be compared to the present and to ask those involved in both past and present situations to comment on the differences between them, and so establish a baseline for comparison.

Components of the study

Survey of teachers' assessment practices

The study was conducted in three phases. The first phase was a survey based on a questionnaire used by Brindley (1989) in a study of assessment practices in the AM EP. Since the issues under investigation in the present study were very similar, it was decided to replicate Brindley's survey and to compare the results.

The survey was designed to investigate the following research questions:

- What are the current classroom assessment practices of the AM EP?
- Do these practices differ from those revealed in the results obtained from the survey undertaken by Brindley in 1989?
- Could different responses from those obtained by the survey conducted by Brindley in 1989 be deemed to constitute evidence for the occurrence of a washback effect on classroom assessment practices caused by the introduc-

3 Methodology

Combining research approaches

This study adopted techniques from both qualitative and quantitative approaches, in a complementary manner. As Goetz and le Compte (1984:245) argue, these should not be seen as mutually exclusive:

Our position is that the transformation of such ideas into dichotomous choices is unnecessary, inaccurate, and ultimately counterproductive.

In order to enhance both the validity and the reliability of this study, different techniques were used to examine the context under investigation. These consisted of a survey, interviews and classroom observations, each providing different but complementary data. This use of multiple data sources is consistent with the notion of methodological triangulation (for example Alderson 1992:285; Alderson and Scott 1992; Cohen and Manion 1989). According to

---

108 Studies in immigrant English language assessment

---

109 The impact of assessment in the Certificate in Spoken and Written English
The impact of assessment in the Certificate in Spoken and Written English

A test will influence what teachers teach;
A test will influence how teachers teach;
Tests will have washback effects for ... some teachers, but not for others. (Alderson and Wall 1993:120-1; author's emphases)

Classroom observations
The third research phase was a series of classroom observations, designed to answer the following research questions:
• Is there an observable difference in the teaching of those who believe their teaching has been greatly influenced by the implementation of the assessment of CSWE and those who do not?
• If teachers report changes to their classroom and other teaching practices as a result of the implementation of the assessment, can any evidence be found through classroom observations to support the view that classroom assessment practices have changed?
• Could such evidence be deemed to constitute evidence for the occurrence of a washback effect?
• Is the washback effect applicable to the context of classroom-based assessment and, if so, to what extent?

4 The survey
Introduction
As noted previously, in order to investigate teachers’ attitudes and beliefs about assessment, it was decided to replicate a survey conducted by Brindley (1989: 167–72). The original survey asked teachers about the usefulness and their use of a variety of teaching and assessment practices. The 1994 survey administered as part of the present study replicated these questions and also included a series of questions concerned specifically with aspects of implementation. It was administered to a total of 215 respondents (see Table 25).

Limitations
Survey population
Although it had been intended to gain as large a pool of respondents as possi-
The low response from Victoria was particularly problematic since this state employed several hundred teachers. Consequently, the survey was neither generalisable nor representative of AMEP teachers nationally and any form of comparison between states became impossible.

The non-random nature of the survey population was also problematic. Where the survey was administered by attending a staff meeting and administering the survey in person, the response rate was very high. Where this was not possible, the surveys were sent to a representative for distribution and, here, the response rate was much lower and self-selected. This, too, severely limited the generalisability of the results.

Survey design

Problems were also found in the design of the survey. First, the questions concerning the percentage of time spent on elements of course design proved difficult for the respondents, as described below. Second, the items which attempted to ascertain the respondents' opinions of the changes showed low (or in some cases, negative) correlations between forms which were intended to be equivalent (Table 26). It would thus appear that in some cases, the practice of constructing ‘parallel forms’ of an item to increase reliability (Reid 1990:328) may result in a loss of validity. In their analysis of a questionnaire they used in a language program evaluation Alderson and Scott (1992:53) reach the same conclusion:

The questionnaires showed signs — perhaps inevitably — of weaknesses, despite piloting. Occasionally questions which were supposed to be repeated ... for purposes of triangulation ... are repeated in different words, which reduces the levels of response overall and dilutes the respondent's opinions over more than one option.

The expected correlation coefficient for these items was ±.85 (Hatch and Lazaraton 1991:441), which clearly was not met. When the offending items were examined, however, it was simple to see, for example, that teaching materials and methods were not the same and that the bald statement of Question 5.11 and the negative in Question 5.6 might be indicators of the problems with these items. The respondents' comments often served to indicate where the problems lay, although,

... one caveat is warranted. The comments provided by teachers were not solicited; the only comments we recorded were volunteered by the teachers surveyed. We do not claim that these comments represent the thinking of all teachers, or even of all the teachers in the study. On the other hand, they do represent the perspectives of teachers who wanted to provide input ... (Cizek et al 1995:169)
The overall results of the 1994 survey, as a percentage of the total responses (n = 215), appear in Table 27, where NR = no response and AG = the Assessment Guidelines.

**Table 27: Number of responses to all survey questions as a percentage of total responses**

<table>
<thead>
<tr>
<th>Question</th>
<th>1 Years in AMES</th>
<th>2 Aspects of teaching</th>
<th>3 Functions of assessment</th>
<th>4 Means of assessing learners’ achievement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&lt;1</td>
<td>1-2</td>
<td>2-5</td>
<td>&gt;5</td>
</tr>
<tr>
<td>1 Years in AMES %</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 Aspects of teaching %</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 = not at all important; no time</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 = extremely important; 50% of course time</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>initial placement: importance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>initial placement: time spent</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pre-course planning: importance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pre-course planning: time spent</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>initial diagnostic assessment: importance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>initial diagnostic assessment: time spent</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ongoing programming: importance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ongoing programming: time spent</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>materials preparation and development: importance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>materials preparation and development: time spent</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ongoing assessment: importance</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ongoing assessment: time spent</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The impact of assessment in the Certificate in Spoken and Written English

**Table 27, continued**

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>NR</th>
</tr>
</thead>
<tbody>
<tr>
<td>final course evaluation: importance</td>
<td>0.47</td>
<td>4.65</td>
<td>6.51</td>
<td>24.19</td>
<td>26.51</td>
<td>36.74</td>
<td>0.93</td>
</tr>
<tr>
<td>final course evaluation: time spent</td>
<td>1.86</td>
<td>28.84</td>
<td>17.21</td>
<td>17.21</td>
<td>11.63</td>
<td>15.81</td>
<td>7.44</td>
</tr>
<tr>
<td>3 Functions of assessment importance %</td>
<td>0 = not important; 5 = extremely important</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>to place students in class</td>
<td>0.47</td>
<td>0.93</td>
<td>0.47</td>
<td>9.77</td>
<td>25.58</td>
<td>62.33</td>
<td>0.47</td>
</tr>
<tr>
<td>to provide feedback on progress to learners</td>
<td>1.4</td>
<td>0.47</td>
<td>2.33</td>
<td>22.33</td>
<td>36.28</td>
<td>36.28</td>
<td>0.93</td>
</tr>
<tr>
<td>to provide information on learners’ strengths and weaknesses</td>
<td>0.93</td>
<td>2.33</td>
<td>4.19</td>
<td>16.28</td>
<td>33.39</td>
<td>42.79</td>
<td>0.93</td>
</tr>
<tr>
<td>to provide information to funding bodies for accountability purposes</td>
<td>11.16</td>
<td>13.49</td>
<td>13.49</td>
<td>26.51</td>
<td>18.14</td>
<td>14.88</td>
<td>2.33</td>
</tr>
<tr>
<td>to encourage students to take responsibility for their own learning</td>
<td>0.93</td>
<td>0.93</td>
<td>6.05</td>
<td>13.49</td>
<td>34.42</td>
<td>43.72</td>
<td>0.93</td>
</tr>
<tr>
<td>to provide students with a record of their own achievement</td>
<td>0.93</td>
<td>2.33</td>
<td>9.77</td>
<td>20</td>
<td>32.09</td>
<td>33.02</td>
<td>1.86</td>
</tr>
<tr>
<td>4 Means of assessing learners’ achievement %</td>
<td>0 = not used at all, completely useless; 5 = very high frequency, very useful</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>observation followed by recycling: frequency</td>
<td>2.33</td>
<td>2.33</td>
<td>6.51</td>
<td>19.53</td>
<td>33.95</td>
<td>33.49</td>
<td>1.86</td>
</tr>
<tr>
<td>observation followed by recycling: usefulness</td>
<td>0.93</td>
<td>2.79</td>
<td>4.65</td>
<td>13.95</td>
<td>33.02</td>
<td>40.93</td>
<td>3.72</td>
</tr>
<tr>
<td>informal discussion with learners about their work: frequency</td>
<td>1.86</td>
<td>6.51</td>
<td>13.95</td>
<td>35.81</td>
<td>23.72</td>
<td>15.81</td>
<td>2.33</td>
</tr>
</tbody>
</table>
Table 27, continued

<table>
<thead>
<tr>
<th>Method</th>
<th>Frequency</th>
<th>Usefulness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Informal discussion with learners about their work: usefulness</td>
<td>1.4</td>
<td>4.19</td>
</tr>
<tr>
<td>Teacher-constructed classroom tests: frequency</td>
<td>12.09</td>
<td>22.79</td>
</tr>
<tr>
<td>Teacher-constructed classroom tests: usefulness</td>
<td>12.09</td>
<td>28.37</td>
</tr>
<tr>
<td>Student self-assessment: frequency</td>
<td>9.77</td>
<td>4.19</td>
</tr>
<tr>
<td>Student self-assessment: usefulness</td>
<td>9.77</td>
<td>15.35</td>
</tr>
<tr>
<td>Teacher journal: frequency</td>
<td>15.35</td>
<td>22.33</td>
</tr>
<tr>
<td>Teacher journal: usefulness</td>
<td>15.35</td>
<td>28.37</td>
</tr>
<tr>
<td>Learner journal: frequency</td>
<td>15.35</td>
<td>15.35</td>
</tr>
<tr>
<td>Learner journal: usefulness</td>
<td>15.35</td>
<td>20.47</td>
</tr>
<tr>
<td>Oral proficiency rating: frequency</td>
<td>4.19</td>
<td>9.3</td>
</tr>
<tr>
<td>Oral proficiency rating: usefulness</td>
<td>4.19</td>
<td>15.81</td>
</tr>
<tr>
<td>Feedback from others outside the classroom: frequency</td>
<td>29.3</td>
<td>15.35</td>
</tr>
<tr>
<td>Feedback from others outside the classroom: usefulness</td>
<td>29.3</td>
<td>14.42</td>
</tr>
<tr>
<td>Standardised published tests: frequency</td>
<td>14.42</td>
<td>8.84</td>
</tr>
<tr>
<td>Standardised published tests: usefulness</td>
<td>14.42</td>
<td>14.88</td>
</tr>
<tr>
<td>Tasks taken directly from the AG: frequency</td>
<td>5.58</td>
<td>14.42</td>
</tr>
<tr>
<td>Tasks taken directly from the AG: usefulness</td>
<td>5.58</td>
<td>14.88</td>
</tr>
<tr>
<td>Teacher-conducted tasks based on the AG: frequency</td>
<td>3.26</td>
<td>8.84</td>
</tr>
<tr>
<td>Teacher-conducted tasks based on the AG: usefulness</td>
<td>3.26</td>
<td>14.88</td>
</tr>
</tbody>
</table>

5 Teachers' comments about the implementation:

<table>
<thead>
<tr>
<th>Agreement</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>N</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strongly disagree</td>
<td>15.35</td>
<td>20.47</td>
<td>31.63</td>
<td>26.51</td>
<td>3.26</td>
<td>2.79</td>
<td>6.51</td>
</tr>
<tr>
<td>Strongly agree</td>
<td>15.35</td>
<td>20.47</td>
<td>31.63</td>
<td>26.51</td>
<td>3.26</td>
<td>2.79</td>
<td>6.51</td>
</tr>
<tr>
<td>I schedule assessment of CSWE when I do my pre-course planning</td>
<td>2.33</td>
<td>2.79</td>
<td>16.74</td>
<td>58.60</td>
<td>17.67</td>
<td>1.86</td>
<td>6.51</td>
</tr>
<tr>
<td>Assessing CSWE makes my classes more formal than they were before CSWE</td>
<td>3.72</td>
<td>6.51</td>
<td>16.28</td>
<td>41.40</td>
<td>26.05</td>
<td>6.05</td>
<td>69.3</td>
</tr>
<tr>
<td>Since the introduction of assessment of CSWE my students are more motivated</td>
<td>17.67</td>
<td>23.72</td>
<td>32.09</td>
<td>15.81</td>
<td>3.72</td>
<td>6.05</td>
<td>6.98</td>
</tr>
<tr>
<td>Since the introduction of assessment of CSWE teaching has become more enjoyable for me</td>
<td>21.86</td>
<td>26.05</td>
<td>36.28</td>
<td>8.84</td>
<td>1.86</td>
<td>6.51</td>
<td>5.12</td>
</tr>
<tr>
<td>Since the introduction of assessment of CSWE I use a more restricted range of materials than I did before</td>
<td>11.63</td>
<td>29.77</td>
<td>16.74</td>
<td>30.23</td>
<td>6.51</td>
<td>5.12</td>
<td>6.05</td>
</tr>
</tbody>
</table>

The impact of assessment in the Certificate in Spoken and Written English
### Table 28: Elements of course design: Perceived importance

<table>
<thead>
<tr>
<th>Element</th>
<th>Perceived importance</th>
<th>Perceived importance</th>
<th>Rank</th>
<th>Rank</th>
<th>p = tobs</th>
</tr>
</thead>
<tbody>
<tr>
<td>needs assessment</td>
<td>Brindley Mean S.D.</td>
<td>Burrows Mean S.D.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pre-course planning</td>
<td>4.39 1.04</td>
<td>4.36 1.22</td>
<td>2</td>
<td>3</td>
<td>n.s. 0.244</td>
</tr>
<tr>
<td>initial diagnostic assessment</td>
<td>3.54 1.36</td>
<td>3.95 1.22</td>
<td>7</td>
<td>6</td>
<td>&lt; 0.01 2.8295</td>
</tr>
<tr>
<td>ongoing programming</td>
<td>4.69 0.67</td>
<td>4.69 0.59</td>
<td>1</td>
<td>1</td>
<td>n.s. 0.85</td>
</tr>
<tr>
<td>ongoing assessment</td>
<td>4.30 0.92</td>
<td>4.43 0.79</td>
<td>3</td>
<td>2</td>
<td>n.s. 1.34</td>
</tr>
<tr>
<td>final evaluation</td>
<td>4.12 0.95</td>
<td>4.20 0.82</td>
<td>4</td>
<td>4</td>
<td>n.s. 0.8</td>
</tr>
<tr>
<td>The assessment tasks reflect the way students learn language</td>
<td>4.19 19.07 32.73 59.53</td>
<td>8.84 6.95</td>
<td>4.65</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 29: Elements of course design: Time spent

<table>
<thead>
<tr>
<th>Element</th>
<th>Perceived importance</th>
<th>Perceived importance</th>
<th>Rank</th>
<th>Rank</th>
<th>p = tobs</th>
</tr>
</thead>
<tbody>
<tr>
<td>needs assessment</td>
<td>Brindley Mean S.D.</td>
<td>Burrows Mean S.D.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pre-course planning</td>
<td>2.72 1.60</td>
<td>1.41 1.57</td>
<td>6</td>
<td>7</td>
<td>&lt;0.001 7.44</td>
</tr>
<tr>
<td>initial diagnostic assessment</td>
<td>2.64 1.40</td>
<td>2.14 1.63</td>
<td>7</td>
<td>5</td>
<td>&lt;0.01 3.023</td>
</tr>
<tr>
<td>ongoing programming</td>
<td>2.84 1.31</td>
<td>2.12 1.37</td>
<td>4</td>
<td>6</td>
<td>&lt;0.001 4.898</td>
</tr>
<tr>
<td>ongoing assessment</td>
<td>4.11 1.14</td>
<td>3.56 1.27</td>
<td>1</td>
<td>2</td>
<td>&lt;0.001 4.1698</td>
</tr>
</tbody>
</table>

Elements of course design

In the first series of questions, respondents were asked to rate the importance and the amount of time they spent on specified elements of course design, using Likert scales. This item was difficult for respondents, with some commenting that they found it very difficult or impossible to calculate the time spent as a percentage, while others chose not to answer. Tables 28 and 29 show the comparison of the results of Brindley’s survey (Brindley 1989:23) and the 1994 survey for elements of course design (n.s. = non-significant result, ie no significant difference between the two groups).
The Spearman rank-order correlation (\( \rho \)) between the responses to Question Two for the two surveys for Perceived importance was significant at 0.929 (\( p = 0.003 \)), while that for Time spent was also significant (0.829, \( p = 0.021 \)), suggesting that the two groups ranked the items in a similar order of importance.

### Functions of assessment

The respondents were then asked to comment on the importance of specified functions of assessment. The comparison of the results of Brindley’s survey (op cit:25) and the 1994 survey appears in Table 30 (in Brindley’s \( n = 131 \); in Burrows’, \( n \) varied from 210 to 215).

<table>
<thead>
<tr>
<th>Table 30: Perceived importance of functions of assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Element</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>place learners in class</td>
</tr>
<tr>
<td>provide feedback on progress</td>
</tr>
<tr>
<td>provide information on learners’ strengths and weaknesses for course planning</td>
</tr>
<tr>
<td>provide information to funding authorities for accountability purposes</td>
</tr>
<tr>
<td>encourage students to take responsibility for their own learning</td>
</tr>
<tr>
<td>provide students with a record of their achievement</td>
</tr>
</tbody>
</table>

The most notable aspect of these results is the consistency between the two surveys. Statistically, the importance of each function of assessment had increased in the second survey, except for provide information on learners’ strengths and weaknesses where a slight decrease had been recorded. This, and the increased
importance given to the element encourage students to take responsibility for their own learning resulted in a change in the ranking of these elements. This might have been the result of considerable discussion within the Australian adult TESOL field about increasing autonomy in adult learners (Hagan et al. 1993; Jackson 1993; Jackson 1994; New South Wales Technical and Further Education Commission 1994) but, at the time of writing, there had been no study of any impact on teaching practice of this discussion.

The results of the application of the t-test provide further evidence that the two sets of respondents were answering the questions in similar ways. The only case where the null hypothesis has been rejected is the last, which might lead one to believe that teachers see more value in providing students with a record of their own achievement than they did five years previously. The rank-order correlation between the responses to Question Three, rho = 0.943, (p = 0.005), suggests the hypothesis that the results were strongly related.

### Means of assessing learner achievement

Respondents were asked to rate the frequency of their use and their perception of the usefulness of specified means of assessing learners’ achievement. The results were again compared to Brindley’s (cf Tables 31 and 32). In the column marked Rank Burrows, the total ranking of all the means of assessing learner achievement are given. This includes items not surveyed by Brindley, because these means of assessing learners’ achievement did not exist at the time his survey was conducted. These are listed below in italics, beside the rank of those items appearing in both surveys.

#### Table 31: Means of assessing learners’ achievement: Frequency of use

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean Brindley</th>
<th>Mean Burrows</th>
<th>S.D. Brindley</th>
<th>S.D. Burrows</th>
<th>Rank Brindley</th>
<th>Rank Burrows</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observation followed by recycling</td>
<td>4</td>
<td>3.843</td>
<td>1.139</td>
<td>1.175</td>
<td>1</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Informal discussion</td>
<td>3.319</td>
<td>3.233</td>
<td>1.227</td>
<td>1.193</td>
<td>2</td>
<td>2 (4)</td>
</tr>
<tr>
<td>Teacher-constructed classroom tests</td>
<td>2.155</td>
<td>3.043</td>
<td>1.478</td>
<td>1.429</td>
<td>5</td>
<td>4 (6)</td>
</tr>
<tr>
<td>Student self-assessment procedures</td>
<td>2.301</td>
<td>2.071</td>
<td>1.535</td>
<td>1.334</td>
<td>4</td>
<td>5 (8)</td>
</tr>
<tr>
<td>Teacher journal</td>
<td>1.802</td>
<td>1.793</td>
<td>1.699</td>
<td>1.778</td>
<td>7</td>
<td>6 (9)</td>
</tr>
<tr>
<td>Oral proficiency rating</td>
<td>3.073</td>
<td>3.150</td>
<td>1.635</td>
<td>1.404</td>
<td>3</td>
<td>3 (5)</td>
</tr>
<tr>
<td>Feedback from outside the classroom</td>
<td>1.945</td>
<td>1.502</td>
<td>1.696</td>
<td>1.446</td>
<td>6</td>
<td>7 (10)</td>
</tr>
<tr>
<td>Standardised published tests</td>
<td>.667</td>
<td>0.841</td>
<td>1.224</td>
<td>1.231</td>
<td>9</td>
<td>9 (12)</td>
</tr>
<tr>
<td>Tasks taken directly from the Assessment Guidelines</td>
<td>NA</td>
<td>2.861</td>
<td>NA</td>
<td>1.419</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Teacher-constructed tasks based on the Assessment Guidelines</td>
<td>NA</td>
<td>3.804</td>
<td>NA</td>
<td>1.158</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Teacher-constructed tasks based on the CSWE and not on the Assessment Guidelines</td>
<td>NA</td>
<td>3.279</td>
<td>NA</td>
<td>1.437</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

#### Table 32: Means of assessing learners’ achievement: Perceived usefulness

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean Brindley</th>
<th>Mean Burrows</th>
<th>S.D. Brindley</th>
<th>S.D. Burrows</th>
<th>Rank Brindley</th>
<th>Rank Burrows</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observation followed by recycling</td>
<td>4.135</td>
<td>4.058</td>
<td>.986</td>
<td>1.087</td>
<td>1</td>
<td>1 (1)</td>
</tr>
<tr>
<td>Informal discussion</td>
<td>3.468</td>
<td>3.538</td>
<td>1.206</td>
<td>1.182</td>
<td>2</td>
<td>2 (3)</td>
</tr>
<tr>
<td>Teacher-constructed classroom tests</td>
<td>2.432</td>
<td>3.216</td>
<td>1.450</td>
<td>1.321</td>
<td>6</td>
<td>3 (5)</td>
</tr>
<tr>
<td>Student self-assessment procedures</td>
<td>2.759</td>
<td>2.510</td>
<td>1.509</td>
<td>1.341</td>
<td>4</td>
<td>6 (9)</td>
</tr>
<tr>
<td>Teacher journal</td>
<td>2.311</td>
<td>2.075</td>
<td>1.764</td>
<td>1.709</td>
<td>7</td>
<td>7 (10)</td>
</tr>
<tr>
<td>Learner journal</td>
<td>1.936</td>
<td>1.979</td>
<td>1.612</td>
<td>1.566</td>
<td>8</td>
<td>8 (11)</td>
</tr>
<tr>
<td>Oral proficiency rating</td>
<td>2.87</td>
<td>3.166</td>
<td>1.553</td>
<td>1.340</td>
<td>3</td>
<td>4 (6)</td>
</tr>
<tr>
<td>Feedback from outside the classroom</td>
<td>2.619</td>
<td>2.648</td>
<td>1.762</td>
<td>1.649</td>
<td>5</td>
<td>5 (8)</td>
</tr>
</tbody>
</table>
The impact of assessment in the Certificate in Spoken and Written English

Table 32, continued

<table>
<thead>
<tr>
<th></th>
<th>Mean Brindley</th>
<th>Mean Burrows</th>
<th>S.D. Brindley</th>
<th>S.D. Burrows</th>
<th>Rank Brindley</th>
<th>Rank Burrows</th>
</tr>
</thead>
<tbody>
<tr>
<td>standardised published tests</td>
<td>.99</td>
<td>1.481</td>
<td>1.334</td>
<td>1.46</td>
<td>9</td>
<td>9 (12)</td>
</tr>
<tr>
<td>tasks taken directly from the Assessment Guidelines</td>
<td>NA</td>
<td>3.068</td>
<td>NA</td>
<td>1.282</td>
<td>NA</td>
<td>(7)</td>
</tr>
<tr>
<td>teacher-constructed tasks based on the Assessment Guidelines</td>
<td>NA</td>
<td>3.845</td>
<td>NA</td>
<td>1.121</td>
<td>NA</td>
<td>(2)</td>
</tr>
<tr>
<td>teacher-constructed tasks based on the CSWE and not on the Assessment Guidelines</td>
<td>NA</td>
<td>3.532</td>
<td>NA</td>
<td>1.208</td>
<td>NA</td>
<td>(4)</td>
</tr>
</tbody>
</table>

In terms of their frequency of use very little had changed in the ranking of the original items. The inclusion of the new items led to a change in the ranking, the second and third ranked items becoming teacher-constructed tasks based on the Assessment Guidelines and teacher-constructed tasks based on the CSWE and not on the Assessment Guidelines. These responses would appear to indicate that teachers had adopted the new methods of assessment, as they were obliged to do (cf Introduction). Alternatively, it could indicate that respondents were aware of what was supposed to be occurring and responded accordingly.

The ranking for perceived usefulness (see Table 32 above), above, showed somewhat more change and this increased with the addition of the new items. Added to this, the standard deviations, together with the respondents’ comments, indicated very different ideas concerning the use and perceived usefulness of teacher journal, oral proficiency rating, feedback from outside the classroom and learner journal.

The validity of applying the t-test to the data from Means of assessing learner achievement is highly questionable, due to the possibility that the new items might have influenced the respondents’ ratings for the original items and it was therefore not applied. The rank-order correlation between those items occurring in both surveys was as follows: for frequency of use, $\rho = 0.967$ ($p = 0.000$), indicating a strong association between the ranks; while for perceived usefulness, $\rho = .883$ ($p = 0.002$). These high correlations suggest that the two groups were answering the questions in a similar way.

The items concerned with the CSWE and its Assessment Guidelines mark the point of departure from Brindley’s survey. These items should also be seen in the light of the instructions given to teachers about the use of the Assessment Guidelines (cf Introduction). Respondents’ comments concerning the quality of the new types of assessment were divided and this recurred in the later questions.

The final question in both surveys asked the respondents for general comments on assessment in the AM EP. In Brindley’s survey (1989:31–4), respondents are reported to have been in favour of assessment generally, and the use of some kind of record in particular, although they expressed concerns about:

- the potential for assessment to be used as a measure of ‘programme effectiveness’
- the potential for assessment to be restrictive
- the dangers of ‘teaching to the test’ (the washback effect, perhaps)
- the need for different types of assessment for different types of learners.

In the 1994 survey, almost one-third of all respondents chose to comment on negative and positive aspects of the implementation. Some saw the assessment as beneficial to students, especially as a means of motivation, since the use of explicit assessment accorded with student views of formal education, although others were concerned with student stress and decreased creativity. Other factors that were mentioned were:

- the question of the quality of the assessment processes, particularly its reliability
- that ‘CSWE assessment has been the result of … accountability, rather than specifically educational factors’
- narrowing of the curriculum
- increased focus and enhanced structure.

A number of issues mentioned by the respondents provided useful input to the next stage of the research (cf Bailey 1997). These could be summarised as follows:
The third group, however, stated in their comments and, by implication in their neutral responses, that they had experienced little or no change and that, for example, other factors influenced their enjoyment of teaching or lack of it.

**Conclusion**

Despite the problems with the survey, the results were important for two main reasons. First, it had been hypothesised that a comparison of the results of the 1989 and 1994 surveys would reveal significant differences, thus providing an initial indication of the occurrence of washback. This was not the case: the survey provided no evidence for the occurrence of washback, because the responses to the two surveys were unexpectedly similar. This similarity, however, was to be useful for the rest of this research, as it was indicative of a similarity in the population samples of the two surveys. The importance of this finding was that were any evidence of changes in teaching to be found in later phases of the study, this would not necessarily have been caused merely by differences in the populations studied.

The second important result was found in the written comments. These were used as the basis for the design of the interview questions, indicating, as they did, areas which needed to be pursued further and with more specific questions. Finally, the responses to the new questions were an early indication of the divergent responses to change which this study was to find overall, encapsulated by the idea that the implementation had affected different people differently (Blewchamp 1994; Alderson and Hamp-Lyons 1996; Shohamy et al 1996; Wall 1996; and Watanabe 1996; Bailey 1997).

**5 Teacher interviews**

**The selection of the participants**

The second phase of the study was a series of interviews. The interview subjects were teachers involved in the survey who had indicated their willingness to be interviewed and additional subjects. Since it was impossible to control the many variables inherent in so diverse a population as AMES teachers, it was decided to attempt to select interview subjects who represented that diversity fully rather than to attempt to find a random sample. Teachers were chosen who had recent or current teaching experience using the assessment of the CSWE and who had been employed for more than two years with an AMES. Because the group who had indicated their willingness to be interviewed did...
not sufficiently represent the diversity of the population, further subjects were contacted and invited to participate in this phase of the study.

Second, the subjects were also chosen to represent the three states (NSW, Victoria and Western Australia) where the CSWE was (by 1995) most in use. As with the survey, availability determined that the greatest number of participants were from NSW (see Table 33), while a ‘balance’ of female and male subjects was chosen to represent the approximate ratio in employment (see Table 34).

<table>
<thead>
<tr>
<th>Table 33: Interview subjects by state</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSW</td>
</tr>
<tr>
<td>22</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 34: Biographical data for interview subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender of interview subjects</td>
</tr>
<tr>
<td>Male</td>
</tr>
<tr>
<td>4</td>
</tr>
</tbody>
</table>

* The majority of these 28 had worked for an AMES for more than 10 years.

After piloting, the choice of subjects was then further refined to include a greater preponderance of teachers who had more than five years experience, since it was these teachers who had had the most to say during the pilot, being able to draw more strongly on their experience before and after the implementation of the assessment of the CSWE.

The respondents were asked about the manner and frequency of their assessment practices; the positive and negative features of the implementation; and the implementation as an example of educational change. The questions also covered specific areas in which the respondents’ teaching might have changed: the structure and organisation of the classroom; changes in methodology and materials; changes in teacher and student behaviour; and other changes they had experienced. The respondents were also asked whether the changes they had spoken about were the direct result of the implementation of the assessment and given the opportunity to add anything they wished to their comments. (The full interview schedule is included at Appendix 1.) The interviews themselves ranged in length from 9 minutes and 9 seconds to 59 minutes and 6 seconds, with the average being about 21 minutes. In order to check the consistency of the interpretation of the data, this section of the study involved other personnel (Burrows 1998:161–4).

The respondents’ comments: summary and discussion

The subjects’ responses were directed only by the questions themselves, with the interviewer further guiding their answers only when clarification or greater detail was needed. Only the major issues of general concern will be examined here and the reader is referred to Burrows (1998) for a detailed analysis and discussion of the interviews.

Certain themes emerged as being important to many, if not all of the respondents. These themes recurred throughout the interviews, being mentioned by a large proportion of the respondents, often more than once. These were also frequently the issues which had been mentioned in the respondents’ written comments in the survey and included both concerns and benefits that the respondents felt were related to the implementation of the assessment of the CSWE. These themes, which are discussed in detail below, were:
- assessment per se, including its formality
- course design
- focus
- communication
- failure
- the assessment of the slow-paced learner
- ‘curriculum or assessment’.

Other common themes, which are also discussed below, were those of stress and time. As the interviews progressed it became clear that some of the respondents had found it difficult to focus on assessment alone, despite the questions asking them to do so and despite the efforts many made. There was clearly an
understandable blurring for some respondents between the implementation of
the curriculum and its assessment (cf Introduction), which is explored at the
end of this chapter.

Assessment
The respondents spoke about issues concerning assessment itself, its reliability,
its standardisation, the need to allow time for assessment within one’s program
and the amount of time it took. Since there had been no summative achieve-
ment assessment before the implementation of the CSWE, having to undertake
assessment was, in itself, a change. A number of respondents stated that the
change in their assessment practices was in the degree of formality it involved.
A distinction was drawn between informal and formal assessment by 13 of the
30 participants and 4 of these defined their perception of that difference.
Informal assessment was described as ‘... an indicator to me in terms of teach-
ing and readiness ...’, using techniques such as peer assessment, portfolios and
observation before setting up a ‘formal testing’ of competency achievement.
These comments recall those made in answer to the questions about elements
of course design in the survey, where a number of respondents wrote comments
about the importance of ‘ongoing assessment’ — techniques which were
described there as ‘informal assessment, mainly carried out through observa-
tion, homework and in-class activities’.

Course design
Course design was also described as ‘programming’ and, most commonly,
‘planning’. This issue was specifically mentioned by 16 of the 30 respondents,
several mentioning it five times or more. Respondents said that their planning
had changed to incorporate the competencies, to ensure that they were taught,
or ‘covered’ and assessed.

In order to examine these claims, it was decided to seek class plans from the
teachers who were observed and compare these to their plans previous to the
implementation. Data on course design prior to the implementation, however,
were not readily available and even those teachers who allowed their classes to
be observed were unwilling to share their course plans with the researcher. For
this reason, a comparison of course plans before and after the implementation
had to be abandoned.

Focus
A number of terms were used by the respondents to express ideas which appear
to be essentially similar and have been grouped together under the heading of
‘focus’. These are clearly related to the issue of planning. Respondents men-
tioned feeling more focused and organised since the implementation of the
CSWE, both in their own teaching and in the program in general, a comment
echoed by the teachers in a study of the implementation of National Curriculum
assessment in the United Kingdom (Gipps et al 1995:145). They spoke of having
more guidance for their course design and of feeling more purposeful in their
ability to set goals for their classes and to achieve those goals. Of the 30 parti-
cipants, 19 spoke about this, most in positive terms but 2 as a feeling of having
to be more organised in order to cope. For the majority of these respondents,
however, there was a feeling of having benefited from the implementation in
this respect.

Communication
Another shared theme was that of changes in the communication that took place
between teachers working together, between students, across NSW AMES,
nationally and, most of all, between teachers and students. The respondents
spoke about giving their students more detailed and better feedback and about
discussing with their students the assessment requirements of the courses and the
assessment itself, including the stakes. A number of respondents stated that they
explained that the stakes of the assessment were not high and used this technique
to reduce their students’ concerns about failure. The discussion of assessment,
for example the performance criteria which were to be met, was seen by the
respondents as being a new aspect of teaching since the implementation.

Failure
Almost half of all respondents mentioned the topic of students failing and for
12 of these 14 this was seen as a negative issue. One respondent, who men-
tioned students failing 19 times during the course of her interview, called it ‘... the
most negative aspect of CSWE ...’. This is not surprising when it is recalled
that, before the implementation of the assessment of the CSWE, no student had
ever failed an AMEP course because it was not possible to do so: before the
introduction of competency-based assessment, a student could not fail a course
because the concept did not exist, at least in teachers’ minds. A student either proceeded to a higher class at the end of a course or stayed at the same level, almost always the former. History does not relate the students’ view of this, however.

Assessment of the slow-paced learner

The other topic which was of concern to half of the respondents, one mentioning it 11 times, was that they did not feel that the curriculum and its assessment were appropriate for all learners and there were some students for whom the assessment system was inappropriate. They identified these students as beginner, or Stage 1 students, the elderly, those who had suffered trauma in their countries of origin, and most commonly Band A (slow-paced) learners.

Curriculum or assessment?

The other general issue raised by respondents was the difficulty of distinguishing between the implementation of the curriculum and its assessment. Many respondents stated frankly that they found it impossible to separate assessment from the other changes which had occurred and this was most clearly the case with the implementation of the CSWE and its assessment. This is an important issue in a washback study, since,

... technically speaking, evidence of teaching and learning effects should be interpreted as washback ... only if that evidence can be linked to the introduction and use of the test. (Messick 1996:243)

Messick goes on:

... it is problematic to claim evidence of test washback if a logical or evidential link cannot be forged between the teaching or learning outcomes and the test properties thought to influence them. (p 247)

Since teaching objectives, learning outcomes and assessment criteria are found in the competencies listed in the curriculum document, it was clearly difficult in this study to identify effects that could be unambiguously attributed to the introduction of CSWE assessment. Every effort was made during the interviews to assist the respondents to separate the assessment from the curriculum when describing the changes they had experienced. It was at this point, however, that the notion of washback began to take on greater complexity, since it became clear that the present study was concerned with connected assessment (Goldstein 1989:140), that is, an assessment system which is related intrinsically to the curriculum for which it was introduced. As criterion-referenced achievement assessment, the assessment of the CSWE was designed to relate closely to what was being taught by the teachers in preparation for the assessment task. This is in contrast to other studies of washback which have been concerned with the effects of ‘curriculum-free’ proficiency testing, or what Goldstein (ibid) refers to as separate assessment.

Although the respondents were asked to distinguish between the assessment and the curriculum, their inability to do so actually reinforced the purpose of achievement assessment in the CSWE:

Achievement assessment measures what a student is able to do as a result of having undertaken language study. In the context of the Certificate in Spoken and Written English achievement assessment refers to what a student is able to do in English in specified situations as a result of language courses undertaken. (NSW AMES 1995:viii)

Since the competencies in the CSWE both guided course design and set the parameters for assessment, it is not surprising that many respondents felt them to be inseparable.

Overarching issues

Three further overarching issues are of importance to the findings of this phase of the study: memory and retrospective research; the changing role of the teacher; and the adoption of the assessment system.

The first issue which must be considered in the context of these interviews was the role of memory in the respondents’ replies to the questions and the validity and dependability of retrospective research. As noted previously, where longitudinal research is impossible, it is sometimes necessary to ask respondents to reflect on past events. In this study, the interview questions asked the respondents to speak about the changes which they might have experienced over a number of years. Interestingly, only two respondents stated that they had had any difficulty in remembering their experiences before the implementation and comparing these to their experiences after this. This is not to say, however, that the respondents’ memories were foolproof nor that time might not have coloured their responses over the intervening years. Moreover, while there is also the consideration of the accuracy of self-reporting: ‘... participants’ reported perceptions of events ...’ (Alderson and Wall 1993:127; my emphasis), the researcher was reliant on the memories of the interviewees. Retrospective research, removed from the present and reliant on memory, is always problematic, since it is not possible to ensure the accuracy of the
information being supplied. This is an inevitable aspect of the ‘... point-of-entry problem ...’ (Beretta 1992:262).

Nevertheless, the interview data appear to be internally consistent: the respondents very rarely contradicted themselves. Further, when questioned at times when they appeared to do so, they were able to explain their responses. It was also apparent that the interviews evidenced a mature reflection on the process of implementation, available to respondents and researcher alike only after time had passed so that in this sense the interview data were more revealing. This reflection can be most clearly seen in the responses of those teachers who said that their initial reaction to the implementation of the assessment, and that of their colleagues, had been very negative but that they no longer felt so strongly, an attitude to change described in Fullan (1991:48) and noted by Shohamy et al (1996). Indeed, several such respondents showed a degree of acceptance of the assessment which went far beyond mere resignation. Amongst these was Teacher B, one of the four subjects chosen for observation, whose responses are discussed in the following section.

The second issue for this phase of the research was the changing role of the teacher (Fitzgerald 1996:8), and particularly the associated factors of teacher stress, paperwork and time. However, there was no group consensus concerning changes in their role, as the comments of these two respondents illustrate:

• They’re really frightened. So it requires a different … treatment from the teacher because if you’re going to … get them to develop their, their abilities, then you have to really take that on board and … nurture them a lot more. I feel like a social worker sometimes, not a teacher.

• … the whole thing is I think a movement away from teachers as carers into teachers as trainers to pass assessments ...

It was decided at the outset of the study that this study would not investigate and discuss the psychological issues surrounding change, nor did the researcher have the expertise to do so, although the concept of teachers’ beliefs will be examined in the final section. Studies of teacher stress, however, have been undertaken by other researchers (Forbes 1973; Smith 1991:9), perhaps partly as a response to the pressures on teachers to change their teaching in the ‘post-modern’ era (Hargreaves 1994), while the negative impact on teacher morale and motivation as a result of ‘... testing mandates’ has been identified in other studies (for example, Corbett and Wilson 1990, in Fullan 1991:24–5). Two studies of teacher stress undertaken in the 1990s (O’Connor and Clarke 1990; Brown and Ralph 1992) recorded time constraints as one of the great causes of stress: ‘The greatest sources of stress were time and work-load pressures’ (O’Connor and Clarke 1990:41). Other studies (Raethel 1996; Spence 1996:5) echoed these results.

Of the 30 interview respondents, 10 mentioned stress, 23 mentioned the time-consuming nature of assessment and 10 respondents mentioned paperwork. However, since these results are congruent with the findings of contemporary studies of teacher stress which suggest that teachers internationally were experiencing the same changes, they cannot be interpreted as evidence for the specific washback effects of CSWE assessment. Nevertheless, the demands of the assessment of the CSWE were clearly perceived as important by many of the respondents.

The third issue was that of the degree to which the assessment system had been adopted. While the responses to the survey indicated that respondents had adopted the assessment, at least in those aspects canvassed therein, three respondents stated that they had only adopted the assessment to the extent that they would, as one put it, ‘... be seen to remain as operating professionally within the system of competency-based teaching in AM E S’. For such a teacher, one would expect there to be little if any impact from the implementation of the assessment, since it could not truly be said to have been implemented. One such teacher was Teacher A, who was one of the four teachers observed during the final phase of the study.

For other respondents, the extent of their adoption of the system would appear to be a combination of adoption and adaptation. A number of the teachers stated that they had adapted the assessment system for their own use in ways that seemed to them to be more practical, more equitable and more realistic. In this regard, Fullan (1991) comments that

... any significant innovation, if it is to result in change, requires individual implementers to work out their own meaning. ... Thus, effective implementation is a process of clarification. ... Clarification is likely to come in large part through practice ... (Fullan 1991:106; author’s emphases)

Wall (1996:351) supported this view, stating that it might take a long time ‘... before an innovation “takes hold” and that there will be many different interpretations of the original idea’.

The interviewees mentioned adaptations to the assessment system. These included:
sometimes happens is that teachers end up fudging a bit … if they think the student is really motivated and doing their best …

• I would like to add that I do at times feel tempted and I know I'm on tape so I won't say that I give way to the temptation but I will admit to being tempted, and you can draw your own conclusions from that, to ignore sometimes the elements of the competencies … so if I’ve got a student that’s … passed the competency apart from that (using the definite and indefinite article correctly) but I know that when they do an exercise in it they get it right, well I pass them and — oh I wasn’t supposed to say that — I’m tempted to pass them …

• … as a teacher I've never really followed the strict guidelines with the time limits because I think every class is different. I mean the class I've got at the moment are … slower English for Work 3 compared to other classes so I feel the students are under enough stress as it is without sort of imposing a really severe time limit …

By the same token, many respondents cited assessment practices which were absolutely congruent with the instructions in the Assessment Guidelines including planning for and timing of the assessments. One teacher described these in the following terms:

… when I feel the students are ready I will be undertaking a formal assessment task which has been validated with other teachers and administer it under the … guidelines, the range statements.

Others mentioned a variety of other practices that were outlined in the Assessment Guidelines, including:

• the use of different types of assessment tasks, both formal and informal, for different purposes
• the use of a second rater to increase reliability, described by a number of respondents throughout the interviews
• the design of assessment tasks based on the model tasks in the Assessment Guidelines and adapted from authentic materials.

In view of the need to ascertain to what extent these self-reported practices corresponded to actual classroom behaviour, however, it was necessary to gather first-hand data from the classroom.
Conclusions and hypotheses for further investigation

The analysis of the interviews led to a number of conclusions and to a number of specific hypotheses which guided the observation phase of the study. Of those conclusions reached, perhaps the single most important was that in this instance of 'connected' assessment it was not only impossible but also unnecessary to divide the impact of the implementation of the curriculum from the impact of the implementation of the assessment, the two being intrinsically interwoven and designed to be so.

Throughout the interviews, the respondents described the many changes they had experienced, the causes of these notwithstanding. When asked to state whether the changes they had spoken about were directly related to the implementation of the assessment per se, the respondents fell into two groups. Almost one-third stated simply that the changes were directly related. Other respondents saw the changes they mentioned as being more related to other causes, particularly the CSWE itself, citing the intrinsic relationship between the curriculum and its assessment.

To examine whether there was an important difference between these two groups, and so to examine the question of causality, it was decided to compare the answers of these two groups to the series of questions which dealt with specific areas of potential change:

- structure and organisation of the classroom
- different teaching methods
- different teaching materials
- other changes in your own behaviour as a teacher
- changes in the students' behaviour
- any (other) changes in the classroom.

This was undertaken to determine whether the changes spoken about by the two groups were different. The question of causality has been identified as a most important aspect of washback research by Messick (1996:243): in this instance, it was necessary to determine whether the implementation of the assessment of the CSWE was the cause of the changes the teachers had experienced or merely simultaneous to the implementation, in order to avoid ‘... the classical post hoc ergo propter hoc fallacy’ (Cohen and Manion 1989:180). There was, in fact, a great similarity between the answers of the two groups which are documented in Table 35.

Table 35: Comparison of changes reported to have occurred since the implementation and changes reported to have been caused by the implementation

<table>
<thead>
<tr>
<th>Changes since the implementation</th>
<th>Changes caused by the assessment implementation of assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assessment related changes</td>
<td>Assessment related changes</td>
</tr>
<tr>
<td>existence of formal assessment</td>
<td>existence of formal assessment</td>
</tr>
<tr>
<td>degree of formality of assessment</td>
<td>degree of formality of assessment</td>
</tr>
<tr>
<td>time management, particularly</td>
<td>time management, particularly</td>
</tr>
<tr>
<td>for assessment</td>
<td>timetabling for assessment</td>
</tr>
<tr>
<td>structuring the courses around</td>
<td>structuring the courses around</td>
</tr>
<tr>
<td>competencies and their assessment</td>
<td>competencies and their assessment</td>
</tr>
<tr>
<td>passing and failing students</td>
<td>passing and failing students</td>
</tr>
<tr>
<td>cheating</td>
<td>cheating</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Teaching related changes</td>
<td>Teaching related changes</td>
</tr>
<tr>
<td>use of authentic teaching materials</td>
<td>use of authentic teaching materials</td>
</tr>
<tr>
<td>use of teaching materials designed by the respondent</td>
<td>use of teaching materials designed by the respondent</td>
</tr>
<tr>
<td>increased concentration on</td>
<td>increased concentration on</td>
</tr>
<tr>
<td>teaching of literacy</td>
<td>teaching of literacy</td>
</tr>
<tr>
<td>n/a</td>
<td>increased accountability and responsibility for teachers</td>
</tr>
<tr>
<td>increased use of meta-language,</td>
<td>increased use of meta-language, inside and outside the</td>
</tr>
<tr>
<td>inside and outside the classroom</td>
<td>classroom</td>
</tr>
<tr>
<td>teaching new topics, particularly</td>
<td>those related to competencies</td>
</tr>
<tr>
<td>those related to competencies</td>
<td></td>
</tr>
<tr>
<td>Teacher–Student related changes</td>
<td>Teacher–Student related changes</td>
</tr>
<tr>
<td>increased student motivation</td>
<td>increased student motivation</td>
</tr>
<tr>
<td>detailed monitoring of student</td>
<td>detailed monitoring of student</td>
</tr>
<tr>
<td>progress and information given to students about this</td>
<td>progress and information given to students about this</td>
</tr>
<tr>
<td>information given to students</td>
<td>information given to students</td>
</tr>
<tr>
<td>about assessment, competencies</td>
<td>information given to students</td>
</tr>
<tr>
<td>and performance criteria</td>
<td>about assessment, competencies and performance criteria</td>
</tr>
</tbody>
</table>

The great similarity of the changes reported by these two groups would seem to make it quite clear that the distinction between the implementation of the assessment and the curriculum framework was simply a false distinction and
Based on the interviews, the following hypotheses were adopted:

- that those teachers who claimed that their teaching had changed would show observable evidence of teaching more literacy, with an emphasis on the teaching of whole texts
- that they would speak to their students about competencies and their assessment
- that they would use teaching materials which supported whole texts and a functional view of grammar
- that they would use more authentic texts; and
- that they would assess their students and prepare their students for assessment related to the assessment of CSE.

These hypotheses were investigated through the observation phase of the study which is reported in the following section.

6 Teacher observations

Models of teacher assessment and response to innovation

The idea of ‘emerging models of teacher assessment’ has been put forward by McCallum et al (1995:57) who described three models or ideal types (op cit pp. 63ff), based on their analyses of teachers’ implementation of National Curriculum assessment in the UK. To explore the concept of teacher-assessors as models, in the present study, four teachers were chosen for observation on the basis of their responses to the interviews: a purposive, maximum variation sample (Merriam 1988:48).

For reasons of practicality, all four teachers chosen were from NSW AMES. Personal information has been limited to preserve their anonymity (cf Table 36).

During the interviews, three of these teachers, Teachers B, C and D, had stated that their teaching had changed, at least partly as the result of the implementation of the assessment, while Teacher A had stated that her teaching was substantially unchanged. It was initially hypothesised, therefore, that the teaching of Teachers B, C and D would be more similar to one another than the teaching of Teacher A. Thus it was hypothesised that, if the teachers’ statements about themselves were correct, there should be observable similarities between the teaching of Teachers B, C and D, other differences in their classes notwithstanding, while the teaching of one which did not need to be taken into account in the final phase of the study.

The second major conclusion that emerged from an examination of the interview data was that change had occurred for many teachers, but not for all, and to differing degrees. It also appeared that, were the respondents’ comments to be accepted, this change may differ with the passage of time, as Shohamy et al (1996) had found. This led to an examination of the comments of those who claimed not to have changed their teaching. In simple terms, these revealed that these teachers felt that the implementation was of benefit neither to themselves nor to their students and had, for this reason, resisted it (Clark 1987). As Fullan and Park (1981:7) state, teacher beliefs ‘... are often critical to effective implementation ... They are also extremely difficult to change.’ In order to explore this phenomenon in greater depth, it was decided to examine the notion of the ‘resister’, the teacher who determines to resist the impact of the implementation of a new assessment system or test. Such a teacher would, it is posited, show few, if any, of the changes mentioned by those teachers who stated that their teaching had changed. On the basis of comments made in the interviews, it was hypothesised that observation would demonstrate that Teacher A was such a teacher, as will be discussed in the following section.
Table 36: Summary of information on observed teachers

<table>
<thead>
<tr>
<th>Teacher A</th>
<th>Teacher B</th>
<th>Teacher C</th>
<th>Teacher D</th>
</tr>
</thead>
<tbody>
<tr>
<td>female</td>
<td>male</td>
<td>female</td>
<td>female</td>
</tr>
<tr>
<td>ESB</td>
<td>ESB</td>
<td>ESB</td>
<td>NESB</td>
</tr>
<tr>
<td>born in</td>
<td>born in</td>
<td>born outside</td>
<td>born outside</td>
</tr>
<tr>
<td>Australia</td>
<td>Australia</td>
<td>Australia</td>
<td>Australia</td>
</tr>
<tr>
<td>had not undertaken</td>
<td>had undertaken</td>
<td>had undertaken</td>
<td>had not undertaken</td>
</tr>
<tr>
<td>post-graduate study</td>
<td>in the TESOL field</td>
<td>study in the TESOL field</td>
<td>in the TESOL field</td>
</tr>
<tr>
<td>TESOL field</td>
<td>field</td>
<td>teaching CSWE II</td>
<td>teaching CSWE II</td>
</tr>
</tbody>
</table>

Teacher A would be observably different in terms of the hypotheses proposed at the end of Section 5.

In effect, if what the teachers had said about themselves was substantially correct and if that translated into classroom behaviours, then Teacher A would act as a type of window into the past, or almost as a 'control'. Teacher A might also be said to be evidence for a new aspect of washback, resistance. While studies have spoken about the notion of resistance to educational change (for example, Burke and Greenglass 1989; Fullan 1991; Markee 1997), Wall (1996:338) may have been the first to have specifically explored the relationship of educational change to washback. Continuing this line of investigation, the present study set out to examine the concept of resistance to washback, particularly in the case of Teacher A.

Because of the intrinsic relationship between the assessment and the curriculum, it was felt that the area of curriculum innovation might have something to offer in attempting to categorise teachers' behaviour in response to changes in curriculum and assessment. Markee (1997:43–4) cited Lambright and Flynn's use of the terms adopters, resisters and implementers (1980, in Markee op cit); and the different types of adopters on a '... diffusion curve: innovators, early adopters, early majority, late majority, and laggards (Huberman 1973; Rogers 1983)' (Markee 1997:58). Bearing in mind the choice of purposive sampling for this phase of the study, the four teachers chosen for observation might be placed on a continuum, ranging from the teacher who felt least affected by the change to the teacher who felt most deeply affected based on their interview comments (see Table 37).

Table 37: Extent of stated changes for observed teachers, teachers as models, from interviews

<table>
<thead>
<tr>
<th>Teacher A</th>
<th>Teacher D</th>
<th>Teacher B</th>
<th>Teacher C</th>
</tr>
</thead>
<tbody>
<tr>
<td>• I don't think the assessment tests meaningful, creative use of the language and that's what I'm interested in.</td>
<td>• ... I'm far more outcome oriented, the broader statement of the competency is equated to the broader aims or objectives of my teaching module.</td>
<td>• ... I've tried not to throw the baby out with the bath water ... the teaching had changed with the passage of time;</td>
<td>• ... the assessment and the teaching are absolutely inseparably linked ...</td>
</tr>
<tr>
<td>• teaching was substantially unchanged;</td>
<td>• teaching had changed in some respects;</td>
<td>• used assessment all the time, subconsciously;</td>
<td>• teaching had changed;</td>
</tr>
<tr>
<td>• had adopted the assessment only as required and where she felt it was relevant;</td>
<td>• had adopted the assessment;</td>
<td>• was accustomed to testing;</td>
<td>• used assessment most of the time;</td>
</tr>
<tr>
<td>• felt curriculum and assessment were interrelated;</td>
<td>• now used a competency-based syllabus;</td>
<td>• fitted the competencies to teaching themes;</td>
<td>• explained the performance criteria to her students;</td>
</tr>
<tr>
<td>• did not find the implementation a confronting or exciting change.</td>
<td>• related outcomes to competencies, performance criteria and underlying skills;</td>
<td>• still used the same methods;</td>
<td>• understood the theoretical bases of the Certificate;</td>
</tr>
<tr>
<td></td>
<td>• liked the genre basis of the competencies;</td>
<td>• implementation allowed one to do virtually the same things if one could see how they fitted.</td>
<td>• had found it a deep change.</td>
</tr>
</tbody>
</table>

Resister adopter (partial) later adopter adopter
The placement of the teachers along the continuum is based on the hypothesis that the teacher who felt least affected by the implementation would show the least observable evidence of washback, while the teacher who felt most affected would show the most evidence. This would mean that Teacher A, who identified few changes to her teaching and saw the change as neither confronting nor exciting, would be unlikely to show evidence of washback in her teaching. Teacher C, at the opposite end of the continuum, mentioned many changes to her teaching and described the implementation as a deep change. She would, therefore, be most likely to show evidence of washback in her teaching. The placement on the continuum of Teachers B and C is less certain, although Teacher B’s description of having moved from resistance to adoption over time seemed to place him nearer to Teacher A than Teacher B. Teacher D did not describe the implementation as a deep change, placing her nearer to Teacher A than Teacher C. The proposition that these four teachers might represent models of teacher behaviour will be examined further in the final section.

The observations

Tools and methods

A number of observation tools (Sinclair and Coulthard 1975; Stiggins and Conklin 1992; Mitchell et al 1981) were examined in terms of their reliability, adaptability and facility of use before the decision was taken to employ the COLT (Communicative Orientation of Language Teaching) Scheme (Spada and Fröhlich 1995). COLT was selected as the most appropriate scheme because it had been shown to be adaptable to a variety of situations, although criticisms of COLT, such as Chaudron (1988) and comments on its limitations (Chaudron et al 1988) were noted. Only COLT A was used (Burrows 1998:314). Two other observers, apart from the researcher, were used for reliability and raters were trained (Burrows 1998:316-17).

The lessons were also observed qualitatively, with the observers recording those times the teachers referred to the CSWE and its assessment, including any concomitant parts (performance criteria, range statements, competencies, credentials). The teachers’ meta-language was also examined, with a focus on whether they used terminology and teaching approaches related to traditional or systemic functional grammar. The teachers were not interviewed about their lessons.

It was determined to observe each of the teachers for two, four-hour lessons, which was felt to be sufficient to allow the teachers to demonstrate a repertoire of teaching techniques, activities and materials and did not over-tax the teachers. Each lesson was recorded on video, with the permission of the students. The reader is referred to Burrows (1998:328–38, and Appendix E) for a description of the content of the eight lessons. It should be noted that while the eight lessons were scheduled to be of the same length, this was not actually the case (cf Table 38). In the following discussion, the letter refers to the teacher and the number to the lesson, so that C2 is Teacher C’s second lesson.

Table 38: Total lesson times in minutes

<table>
<thead>
<tr>
<th></th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
<th>C1</th>
<th>C2</th>
<th>D1</th>
<th>D2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>194m</td>
<td>208.25m</td>
<td>212.75m</td>
<td>177m</td>
<td>174.25m</td>
<td>193.25</td>
<td>201.75m</td>
<td>199.25m</td>
</tr>
</tbody>
</table>

The researcher found that videotaping had advantages as well as the obvious disadvantages: while not allowing for the whole of the classroom to be seen, and while the quality of sound was not as good as that received by the human ear in situ, the advantages of permitting cross-checking and repeated playing, and thus enhancing inter- and intra-rater reliability, more than compensated for these disadvantages. This may explain, at least in part, the growing use of videotaping in classroom observation (Cohen and Manion 1989:143; Kellehear 1993:132; Spada and Fröhlich 1995:119).

When COLT A is used, each lesson is divided into activities and episodes, which are timed. COLT A then allows for the time-coding of various lesson features under five major sections: Participant organisation, Content, Content control, Student modality and Materials. A detailed explanation of each of these sections may be found in Spada and Fröhlich (1995:13–27).

Hypotheses

Based on the categories within COLT A and the teachers’ interviews, the following hypotheses were proposed:

- Differences in Participant organisation (how the class was organised) and Content control (who was leading the lesson) were not expected, since all four teachers had said that they had tried not to change their teaching in these areas.
The teaching of Teacher A was expected to differ from that of the other three teachers in terms of Content (the theme of activities). It was expected that Teacher A would base her teaching soundly in traditional grammar, with little or no reference to generic whole texts and the structure of these texts. This would be seen in a lower proportion of teaching of Function, Discourse and Sociolinguistics in her lessons and a higher concentration on Form, in comparison with the lessons of the other three teachers. The teaching of Teachers B and C was expected to be similar, while the extent to which Teacher D had adopted the CSWE and its assessment might also be revealed here.

It was expected that Teachers B, C and D would use a greater proportion of extended texts than Teacher A, because of the teaching of generic, whole texts, particularly in the teaching of literacy.

It was expected that Teachers B, C and D, particularly B and C, would spend more lesson time on activities related to literacy.

In terms of materials, Teachers B, C and D were expected to use a greater proportion of authentic texts, native speaker texts and native-speaker adapted texts than Teacher A.

Teachers B, C and D were expected to use more role-plays related to particular spoken language competencies as preparation for assessment tasks, than Teacher A would do.

Teachers B and C were expected to use the meta-language of systemic, functional grammar, while Teacher A was expected to use only the meta-language of traditional grammar. The extent to which Teacher D used the meta-language of systemic, functional grammar might demonstrate the extent of her adoption of the assessment of the CSWE in this respect. While a teacher could use the meta-language of systemic-functional linguistics without having adopted the CSWE and its assessment, it would be extremely difficult to prepare the students for assessment and give feedback on it without using any of this meta-language.

Teachers B, C and D were expected to show more evidence of the use of competency-based assessment than Teacher A. These three teachers were expected to speak with their students about the CSWE and its competencies, performance criteria and range statements, and about assessment tasks. It was further expected to observe evidence of either or both of actual assessment tasks and practice assessment tasks in the teaching of Teachers B, C and also D, who stated emphatically in her interview that she had adopted the assessment system.

### Table 39: Content of lessons observed

<table>
<thead>
<tr>
<th>Content</th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
<th>C1</th>
<th>C2</th>
<th>D1</th>
<th>D2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Procedure %</td>
<td>18.72</td>
<td>3.11</td>
<td>5.05</td>
<td>4.35</td>
<td>1.18</td>
<td>4.34</td>
<td>18.55</td>
<td>13.23</td>
</tr>
<tr>
<td>Form %</td>
<td>76.78</td>
<td>86.67</td>
<td>62.63</td>
<td>5.68</td>
<td>20.88</td>
<td>32.23</td>
<td>70.55</td>
<td>28.66</td>
</tr>
<tr>
<td>Function %</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Discourse %</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Sociolinguistics %</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4.11</td>
<td>0</td>
<td>2.17</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Form + Function %</td>
<td>4.51</td>
<td>5.48</td>
<td>15.4</td>
<td>15.48</td>
<td>0</td>
<td>6.94</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Form + Discourse %</td>
<td>0</td>
<td>0</td>
<td>2.02</td>
<td>0</td>
<td>25.29</td>
<td>15.03</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Form + Sociolinguistics %</td>
<td>0</td>
<td>4.73</td>
<td>0</td>
<td>27.81</td>
<td>9.71</td>
<td>7.51</td>
<td>4.73</td>
<td>18.03</td>
</tr>
<tr>
<td>Form + Function + Sociolinguistics %</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11.61</td>
<td>8.53</td>
<td>0</td>
<td>6.18</td>
<td>3.11</td>
</tr>
<tr>
<td>Form + Discourse + Sociolinguistics %</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>14.16</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Form + Function + Discourse + Sociolinguistics %</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>14.89</td>
<td>27.33</td>
<td>34.41</td>
<td>12.43</td>
<td>0</td>
</tr>
<tr>
<td>Function + Sociolinguistics %</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3.63</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Discourse + Sociolinguistics %</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5.2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Content Total %</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>
As predicted, the Content of Teacher A’s lessons comprised Form and little else. Also as predicted, Teacher C’s lessons were composed of a wide spread of Content areas. Teacher D’s lessons also show a spread across Content areas, although less broad than Teacher C’s. The influence of a one-hour assessment task on the Content in Lesson D2 is clear. Teacher B’s lessons, while also showing a broad spread across Content areas, like Teacher C’s, were more different from one another than those of the other teachers. The reason for this may be found in the topics of the lessons. In the first lesson, Teacher B spent a great deal of the lesson time working with the students on new vocabulary and its pronunciation. He spent a great deal of the second lesson, however, on the genre of the casual conversation and its sociolinguistic place, and also looked at the Function and Discourse associated with it.

The differences in teachers’ practices become easier to see when the Form is broken down into its constituent parts, described in Spada and Fröhlich (1995:37–8) as grammar, vocabulary, pronunciation and spelling. In the following table, the Form aspects of each lesson are broken down as percentages of total Form content. It should be remembered that the number of minutes spent on Form only varies greatly from lesson to lesson.

Table 40: Sub-categories of Form as percentages of Form only

<table>
<thead>
<tr>
<th>Form: Content</th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
<th>C1</th>
<th>C2</th>
<th>D1</th>
<th>D2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar %</td>
<td>62.08</td>
<td>32.47</td>
<td>16.13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Vocabulary %</td>
<td>1.13</td>
<td>9.91</td>
<td>23.59</td>
<td>21.28</td>
<td>64.08</td>
<td>71.3</td>
<td>72.68</td>
<td>45.7</td>
</tr>
<tr>
<td>Grammar + Vocab %</td>
<td>0</td>
<td>38.79</td>
<td>14.11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Grammar + Vocab + Spelling %</td>
<td>0</td>
<td>0</td>
<td>13.31</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Vocab + Spelling %</td>
<td>0</td>
<td>0</td>
<td>10.89</td>
<td>0</td>
<td>8.45</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Vocab + Pronoun %</td>
<td>0</td>
<td>0</td>
<td>21.98</td>
<td>17.02</td>
<td>27.46</td>
<td>20.63</td>
<td>27.32</td>
<td>38.91</td>
</tr>
<tr>
<td>Pronunciation %</td>
<td>0</td>
<td>0</td>
<td>61.7</td>
<td>0</td>
<td>8.07</td>
<td>0</td>
<td>0</td>
<td>15.38</td>
</tr>
<tr>
<td>Pronun + Grammar %</td>
<td>19.19</td>
<td>0.57</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Pronun + Grammar + Vocab %</td>
<td>17.61</td>
<td>11.35</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Pronun + Grammar + Spelling %</td>
<td>0</td>
<td>6.9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total Form Only %</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Total Form Only in minutes</td>
<td>110.75</td>
<td>174</td>
<td>124</td>
<td>11.75</td>
<td>35.5</td>
<td>55.75</td>
<td>97</td>
<td>55.25</td>
</tr>
</tbody>
</table>

Table 40 shows those sections of the lessons which were related to the Form category only. The data above demonstrate that the Content of Teacher A’s lessons is different from that of the other teachers, her concentration on traditional grammar being most evident: almost all of the Content of both of her lessons is related to traditional grammar. There is a notable difference between the lessons of Teacher A: in her first lesson she concentrated on Grammar alone, with some work on Grammar and Pronunciation, while in her second, she worked on Grammar in conjunction with Vocabulary. Teacher A spent the greatest time of any teacher on traditional grammar and demonstrated limited teaching of Function, Discourse and Sociolinguistics, despite the fact that success in CSWE demands that students are proficient in the functional, discourse structure and socio-linguistic aspects of the generic texts specified in the competencies. This would seem to indicate that her lessons are not related to those performance criteria which students must achieve in assessment of the CSWE, providing support for the hypothesis that her lessons would not show evidence of washback. This will be discussed further in the section on meta-language, below.

The Content of the lessons of Teachers C and D was not related to traditional grammar, that is, sentence-level grammar divorced from whole texts, an indication that both would seem to have adopted a systemic-functional perspective, as they claimed to have done in their interviews. The adoption of the Teaching/Learning Cycle (Joyce and Burns 1992, in Cornish 1992:17–18) was an example of change in teaching methods mentioned by a number of the interviewees and provided support for the hypothesis that washback would have occurred in this respect. The concentration on Vocabulary in the teaching of Teachers B, C and D could be described as Building the Field, the first stage of the Teaching/Learning cycle, which occurred when these teachers assisted their students to develop the vocabulary they would need to read, listen to, speak and write specific texts.

The biggest difference of all, however, was between the two lessons of Teacher B, in the time spent on Form alone. An examination of Table 40 shows the effect on Form of his teaching written language in the first lesson and spoken language in the second, a feature which recurred throughout the analysis. Lesson B1 is situated between those of Teacher A and those of Teachers C and D, in that he included some traditional grammar and an exercise on active and passive voice. Unlike Teacher A, as will be described in more detail below, he related this grammar to a generic text, the Report. In B2, the main teaching point was the casual conversation genre. This lesson is more like those of
Teachers C and D than Lesson B1. In this lesson, Teacher B spent very little time on Form only. He related his teaching to the genre of the casual conversation, so spending more time on other Content features in combination with Form.

Teacher B was the teacher who claimed to have kept those features of teaching and assessing which pre-dated the CSWE which he still found useful, and to have adopted those features of the CSWE and its assessment for which he had a place. Both of Teacher B’s lessons show evidence of this, since both contain segments when he was teaching Form, Function, Discourse and Sociolinguistics (cf Table 39) which, taught in combination, would appear to demonstrate his use of systemic functional grammar, and both contain segments when he was teaching Form only (cf Table 40). This provides evidence for the hypothesis that the teachers’ descriptions of their teaching would be supported through the observations and for the hypothesis that Teacher B would be observed to be an adopter, albeit a late adopter.

It was then decided to combine the Content of the two lessons of each of the teachers, in order to make an overall comparison. This is common practice in washback studies where observation data is employed (for example Blewchamp 1994; Alderson and Hamp-Lyons 1996; Watanabe 1996), probably because most washback studies observe a number of lessons of shorter duration than those in this study (see Figure 1).

When the teachers’ two lessons are combined, the differences between the four teachers become more apparent, supporting the hypothesis that the lesson Content of Teacher A would be different from that of the other three teachers, with Teachers B and C being similar to one another and Teacher D lying between the two extremes. This, in turn, provides support for the notion that there has been a washback effect on this aspect of the teaching of Teachers B and C and that there has been no effect on this aspect of the teaching of Teacher A. The effect on Teacher D appears to have been more limited.

As described above, all of the teachers had spent a considerable amount of lesson time on Form only and this was an unexpected result. To examine this more closely, the sub-categories of Form have been separated (as in Table 40) and the two lessons of each teacher combined (Figure 2).

When each of the teachers’ lessons are combined, two aspects of the observations become apparent. The first is that Teacher A’s lessons are unlike any of the other three teachers: she alone spent a considerable amount of time teaching traditional grammar both separately and in combination with the other sub-categories covered by Form. In this respect, Teacher B is somewhat more like Teacher A than the other two teachers, although this should be seen in the context of the much smaller percentage of time Teacher B spent on Form altogether than Teacher A and the discussion, above.

Teachers C and D are more like one another in Form in their concentration on vocabulary and pronunciation and in the fact that they spent no time on teaching traditional grammar. There are two possible explanations for this. The first is that these two teachers were teaching higher level students who did not need to study traditional grammar. The second is that both have adopted the CSWE and its assessment in this respect, the more likely explanation when the
this research, ‘at some times and not at others’, ‘to a greater or lesser extent’ and ‘in some respects but not in all respects, for the same teacher’. These modifications will be discussed further in the final section.

Student modality

The four teachers’ lessons were also compared in terms of their Student Modality, that is, the time spent on the four macro skills both alone and in combination, to investigate the hypothesis that Teachers B, C and D would spend more time in their lessons addressing literacy than Teacher A did. Initially, the greatest difference found was between Lessons B1 and B2, again understandable in the context of the topics taught, although when Reading and Writing were isolated from the other macroskills, the greatest difference was between D1 and D2, due to the one-hour written assessment task.

In order to investigate this further, it was necessary to recall the interview comments made about the teaching of literacy and characteristics of the CSWE. The view of literacy which underpinned the CSWE was based in systemic-functional grammar and the competencies themselves were described in terms of generic whole texts. Both Teachers B and C referred to their use of generic texts in teaching. The strong feeling of the interview respondents was that the teaching of literacy had increased as a result of the implementation of the CSWE and its assessment. This might be because the teachers’ concept of what teaching literacy means had changed. Teachers A and D presented a great deal of written material to their students which the students were expected to be able to read. With the exception of the assessment task in Lesson D2, the students in both these classes were expected to undertake written exercises which involved writing only individual words, phrases and sentences.

The difference in the teaching of literacy between Teachers A and D and Teachers B and C might best be described by relating it to the CSWE and its assessment, as follows. Teachers A and D used activities which required their students to read and write but used no explicit means of teaching literacy skills. Teachers B and C used activities in which they taught their students to read and write. In order to prepare their students for the assessment of reading and writing competencies, Teacher B and C used such techniques as deconstruction of texts and those included in the Teaching/Learning cycle (op cit):

- building the context
- modelling the text
difficult to see how teaching and learning of whole texts occurs. Since it is necessary for students to be able to speak, read and write extended texts in order to be successful in the assessment tasks, the findings here support the hypothesis that Teacher A had resisted any change in this aspect of her teaching and had, therefore, by extension, resisted the implementation of the curriculum and its assessment. These findings also support the hypotheses concerning the teaching of Teachers C and D, that Teacher C’s teaching practices would show evidence of washback and that Teacher B would be affected, without abandoning previous practices. In the case of Teacher D, these results are indicative of an area in which the implementation of the assessment has had little impact on her teaching, also consistent with the views expressed in her interview. Finally, both Teachers C and B taught their students (explicitly or by demonstration) the differences between spoken and written language. An understanding of

joint construction of the text
independent construction of the text; and
linking to related texts.

If one accepts that the deconstruction and reconstruction of texts assists students in both reading and writing, then it can be said that Teachers B and C spent a considerable part of their lessons teaching literacy, that is, teaching the functional, discourse structure and sociolinguistic characteristics of the texts they expected the students to read and write. This, then, would demonstrate the difference between the two pairs of teachers, indicating that neither Teacher A nor Teacher D had changed her teaching in this respect. The evidence for these changes in the teaching of Teachers B and C may be found in the Content of their lessons, in which Form, Function, Discourse and Sociolinguistics, together and in various combinations, were taught consistently by these two teachers, while all four teachers were also observed to teach the macroskills in combination (Burrows 1998:360–64). This, too, provided support for the hypothesis that washback had impacted on the teaching of Teachers B and C and that Teacher A had again resisted the implementation and, thus, had not shown evidence of any washback effect. Teacher D would not seem to have been affected in this aspect of her teaching, except in the brief observed preparation of her students for the assessment task.

The differences in the teaching of literacy become clearer when the two lessons of each teacher are combined in order to examine their overall use of minimal and extended texts in the teaching of reading and writing. This led to a comparison of the types of text, minimal or extended, which the four teachers used (see Figure 3). An examination of the graph in Figure 3 shows that, once again, Teacher A and Teacher C are most different from one another. Teacher B is more like Teacher C and, in this case, Teacher D is more like Teacher A. Both Teachers A and D show a much greater use of minimal texts for reading and writing in the lessons observed. Teacher C shows a distinct use of extended texts in her teaching and Teacher B uses minimal and extended texts almost equally and minimal and extended texts more than the other teachers. These results are consistent with the findings related to the four teachers’ teaching of Content: without the use of extended texts, it is difficult to teach grammar above the level of the clause. If a teacher only includes activities in which students listen to, speak, read and, particularly, write words, phrases and sentences, it is
these differences is important in the teaching of the different genres in the CSWE and is of assistance to students undertaking its assessment. The incorporation of this element into their teaching practices also provides some evidence for washback in the lessons of the two teachers.

**Meta-language**

Two other areas were examined in the observations: teachers’ meta-language and their use of and reference to assessment. Analysis using COLT had already provided support for the hypothesis that Teachers B and C would use the meta-language of systemic-functional grammar, while Teacher A would use only the meta-language of traditional grammar. The extent and kind of the meta-language used by Teacher D was not able to be predicted from her interview with confidence, since she only said that she was using ‘... the systemic model more and more’. She did not mention speaking to the students about genre in her interview but did mention speaking to them about the performance criteria, descriptions of the elements of the generic texts represented by the competencies. COLT A does not allow for the coding and therefore measuring of the use of meta-language, so description was employed.

Teacher D spoke specifically about the performance criteria and range statements in the competency on which she assessed her students, using the terminology of both traditional grammar (eg pronouns) and of functional grammar (eg reference). Teacher D also referred to the principle in the Assessment Guidelines, described above, which stated that students should be given more than one attempt to pass each competency.

Teacher A used exclusively the meta-language of traditional grammar. She referred to nouns, verbs, adverbs and adjectives and to the past tense and did not explain any of the terminology she used to her students. She did not describe the use of these parts of speech in terms of their function, nor did she relate their use to any generic text type. She taught grammar only at the sentence level.

In his interview, Teacher B said that he used grammatical meta-language more than he had done previously, specifically mentioning generic text structure and systemic-functional grammar. He was observed teaching three genres: the report, the narrative and the casual conversation. He explained the difference between the report and the narrative, referring to past and present tense and the use of general versus specific nouns and pronouns, implying but not using ‘reference’. When he began work on the report he used the terms ‘paragraph’ and ‘sentence’, later explaining the need for a heading. He taught the structure of the report and explained the function of different types of ‘conjunction’ between clauses and between sentences and also explained textual conjunction. He also explained the difference between ‘active’ and ‘passive’ and the use of the passive voice.

Teacher C used the widest variety of functional grammar meta-language, although she also used traditional grammar terminology. She taught the students about the different types of processes, nominal groups, generic reference, participants, conjunction, tense, text structure and layout. Some of these terms were new to the students and she explained these; others were terms which she reminded them she had taught before, such as nominal groups; and some, such as tense and layout, were assumed knowledge. She also compared different text types. Teacher C spoke about ‘Theme’, the generic structure of a job application letter, and the Sociolinguistic and Functional conventions of a job application letter. She used the terms ‘noun’ and ‘verb’, however, in a pronunciation activity.

Once again, the four teachers behaved differently. Teacher A did not use the meta-language of systemic-functional linguistics at all, another aspect of her teaching which supported her view that her teaching had not changed to any great extent. Teacher D used different terminology for different purposes, using the meta-language of traditional grammar rather than functional grammar, except where functional grammar was necessary to explain the assessment criteria. Teacher C both taught and used the meta-language of functional grammar, relating this to the text-types she was teaching, although she also used traditional grammar terminology. Teacher B taught both traditional and functional grammar, mostly using examples rather than a great deal of meta-language, except where such terminology was necessary, having said in his interview:

> ... I’m not as self conscious about using meta-language now. I came from an RSA Certificate originally and we were taught you do not teach meta-language ... we used to be threatened with failure if we even said noun ...

Each of these examples could be explained as being related to each teacher’s view of grammar and possibly, therefore, as being unrelated to the implementation of the CSWE and its assessment, were it not for the links between the grammar and the assessment.
Assessment

The ‘connectedness’ of the assessment meant that when the teachers were talking about assessment they were inevitably also talking about the curriculum. However, some aspects were seen to be more closely related to the assessment than the curriculum. Consequently, the raters were asked to note instances of the following aspects of teacher behaviour which related specifically to assessment:

a) where the teachers spoke to the students about:
   - the CSWE level of the class
   - achieving a CSWE credential
   - competency-based assessment tasks from the CSWE
   - performance criteria or range statements from the CSWE

b) where the teacher taught:
   - materials which the teacher explicitly stated were related to a competency
   - activities which the teacher described as practice CSWE assessment tasks

   and

c) where the teacher actually assessed a CSWE competency.

There were no examples of these features in Teacher A’s classes. She gave one ‘test’ to her students, which she helped the students complete, and which was not related to any competency.

Teacher B made many references to the assessment of the CSWE. His course was nearer its end than those of the other teachers: according to the interviewees, including Teacher B himself, this situation required planning teaching around the assessment in order to enable the students to achieve Certificate II before the course ended. In his lessons, he referred to the competencies ‘Reading a narrative’ and ‘Writing a report’, explaining when each would be assessed. He began his second lesson by listing the competencies which the students had already achieved in each macro-skill and then indicating the next competency he would be addressing, the casual conversation. The next part of this lesson, two hours, was spent on the features of the casual conversation, with Teacher B showing his students model texts on video, having them practise in pairs and groups and tape recording two pairs of students for peer-assessment. He also used an overhead transparency of the performance criteria, spoke to the students about the range statements and told one of the pairs they would have passed if the activity had been an actual assessment task.

Teacher C had a list of the CSWE competencies on the wall of her classroom and referred to it in both lessons. In the first lesson, she announced that the class would be working on the competency ‘Writing a report’ and proceeded to spend several episodes on this, including the deconstruction of model reports, eliciting the language and structure needed for a report from the students, to make a performance criteria list.

Although Teacher D’s first lesson contained none of the identified items, she administered an actual assessment task in her second lesson, the only teacher observed to do so. In this lesson, she prepared the students by reminding them of the performance criteria and range statements, which she told them they had studied with her the previous day (a lesson which was not observed). The students then had one hour in which to write the report which was observed to take place under examination conditions.

Teachers B, C and D all referred specifically to the CSWE and its assessment but Teacher A did not. All three were explicit in talking to their students about the assessment requirements of the CSWE and all made an effort not to engender a nervous response to the idea of assessment from their students. From the lessons observed, it was clear that both Teachers B and C had incorporated the assessment of the CSWE into their teaching. Teacher D, on the other hand, appeared to have overlaid the assessment onto her lessons in a less connected manner.

The hypothesis that Teachers B, C and D would show more evidence of using the assessment of the CSWE than Teacher A was substantiated by the observations, although Teacher D appeared to have added assessment on to her lessons rather than to have incorporated it into her teaching as a whole. Analysis of assessment-related teaching in the lessons observed indicated a washback effect on the teaching of Teachers B and C, and on Teacher D to a lesser extent. However, it was difficult to see any relationship between the observed teaching of Teacher A and the assessment of the CSWE: the evidence from the observations did not support her statement in her interview when she said she used ‘... competency-based assessment ... Where I think it’s relevant and fits into my scheme of teaching ...’
Summary and conclusion

The four teachers were selected for the observation phase of this study as representative of different models of teacher behaviour, and, of course, on their willingness to participate. By selecting only four teachers, and observing only two lessons of each one, despite the length of those lessons, the researcher chose to limit the generalisability of the results in favour of a purposive choice of subjects. The selection of the sample also decides the types of analysis that might be used, since the use of this model would have made the application of statistical methods invalid. For this reason, only descriptive analysis was employed in this phase of the study.

As hypothesised, the Content of Teacher A’s lessons did, indeed, differ from that of the other three teachers, with a far smaller percentage of teaching time spent on her lessons on Function, Discourse and Sociolinguistics and a far higher concentration on Form. Teacher D fell between the two extremes, supporting her claim that she had adopted the implementation to some extent but not entirely. When Form was broken down into its component parts, Teacher A was found to be most different from Teachers C and D in this respect, with Teacher B using the greatest variety of topics within this category. As hypothesised, the teaching of Teachers B and C was similar in many respects, with the added complication of the difference between the two lessons of Teacher B, supporting his statement that he had not thrown ‘… the baby out with the bath water’, since he taught both traditional and systemic functional grammar. The hypothesis that Teachers B, C and D would spend more time on the teaching of literacy than Teacher A was supported for Teachers B and C but not D, when the concept of the teaching of literacy and the use of extended texts were included. The hypothesis that Teachers B, C and D would spend more time using and teaching extended, generic texts than Teacher A was supported for Teachers B and C, while Teacher D’s use of extended texts lay between these two. Teacher B used both kinds of texts.

The hypothesis that the meta-language of the teachers would be different was also supported. Teachers B and C used more systemic-functional meta-language, particularly in relation to text structure, with Teacher C being the most different from Teacher A. Apart from the section in which she prepared the students for the assessment task, however, Teacher D did not demonstrate an adoption of the meta-language of systemic-functional linguistics.

In terms of the assessment of the CSWE, Teacher A’s lessons showed no evidence of the existence of the assessment tasks nor of the CSWE and its competencies, while the lessons of all of the other three teachers showed evidence of either actual assessment or preparation for this. Teachers B and C had interwoven the assessment into their lessons and Teacher D had included it rather as an adjunct: it could be said that washback was observed in the teaching of Teacher D only as it related specifically to assessment and the observation of only two lessons did not allow for a good understanding of her stated linking of competencies and objectives.

When these results are related back to the washback hypothesis which states that the implementation of assessment will have an impact on teaching practice, the findings of this phase of the research provide support for the idea that the reality, at least in relation to classroom-based assessment, is more complex than this. As has been found in studies of curriculum innovation, and in recent washback studies cited above, the reaction of individual teachers to the implementation of assessment is itself individual. The observations undertaken here have revealed a cline from Teacher A, upon whose teaching the assessment would appear to have had no observable washback effect, to Teacher C, whose teaching demonstrates many of the changes spoken about by the interview subjects. Between these two extremes lie Teachers B and D, upon whose teaching there would appear to have been a washback effect in some respects but not in others. Teacher D would seem to have changed in respect of the Content of her lessons, while in other respects, such as the teaching of whole texts, she would not seem to have changed. The results with regard to Teacher B were still more complex and only revealed by the separate analysis of his two lessons. Teacher B would appear from the observations to have combined those aspects of his lessons which were still applicable in the new context, such as the teaching of traditional grammar, with those aspects which he had decided to adopt, such as the teaching of whole texts and their structures. The result of this in the analysis was a marked distinction between his two lessons and a wider spread across the different sub-categories chosen for analysis. In terms of the inclusion of the assessment, Teacher B showed evidence of a marked washback effect.

These results both confirm and do not support the models advanced at the beginning of this section: the analysis supports the categorisation of Teacher A as a resister, Teacher C as an adopter and Teacher D as a partial adopter. Teacher B, however, would seem to be an ‘adopter’, among those who take from the new system as they choose. Teachers A and B seem to demonstrate an
idea which has been absent from previous washback literature, that of choice. The literature on washback predating Alderson and Wall (1993) described the washback effect as an inevitability. Bearing in mind the limitations posed by the selection and extent of the data, and the complicating factor of the connectedness of the assessment and the curriculum, the results here would appear to provide support for the notion that there is a degree of choice involved in washback, at least in the context of classroom-based assessment. This may be explained as follows. If it is possible to choose to resist the effect of an implementation upon one's teaching, then it is possible to choose whether the implementation of an assessment will have a washback effect. This will be the subject of the final chapter.

7 Summary and conclusion

Introduction

Unlike other studies of washback in the field of applied linguistics, which usually involve large-scale, high-stakes testing, this research concentrated on classroom-based assessment. Although in the course of the present study the assessment of the CSWE became increasingly large-scale and results of teacher assessments were increasingly used to report outcomes, it has neither the stakes nor the degree of standardisation of high profile public tests such as the TOEFL.

This research was also unlike studies of washback in testing because of the connectedness of the assessment and the curriculum, with many of the interview subjects rejecting the distinction between the CSWE and its assessment. This led the researcher to the realisation that such a distinction was unnecessary but, as with the issue of the stakes of the assessment, this lessened the comparability of this research with other washback studies. The problems encountered and the purposive sampling used, with a concomitant decrease in generalisability, should be recalled in the following discussion.

Washback studies and new definitions of washback

In an ideal world, an investigation of washback effects should be set up as an empirical study, with an examination of the chosen educational context, either undertaken before and after the implementation of a specific test (for example Hughes 1989), or using a situation where different teaching situations can be compared (Alderson and Hamp-Lyons 1996, for example). In the world of education, however, such situations do not always arise and, as in the case of the Sri Lankan Impact Study (Alderson and Wall 1992), one must adapt one's methodology to suit the context.

Shohamy et al (1996) argued for a more complex version of washback, where a number of external variables, including test stakes, purpose and time, affected the extent of the impact of a test. The interview segment of this study indicated that a number of the respondents felt that their perception of the implementation of the assessment had altered with time and the discovery that the assessment system represented a smaller change than they had expected: ‘... like a different set of clothing ... clothing the same body ... more of a difference in perception than a fundamental change', as one said.

Watanabe (1996) and Blewchamp (1994) found different results for the different teachers they studied, as did Wall, who reported:

One of the interesting questions which arose during our study was why the exam was having so much impact on some aspects of teaching but so little on others ... There seemed to be many factors influencing the way that teachers handled their classrooms, ranging from their own personal beliefs, to the requirements of their own institutions, to the large-scale economic, social and political issues prevailing at the time. While some teachers found it easy to adjust to the new demands, others found it difficult even to understand, much less go along with, the ideas underlying the new teaching materials and assessment procedures. (Wall 1996:348)

Wall listed a number of factors ‘... which might have prevented the examination from providing an effective “lever for change”' (ibid:348–50). Several of these factors were similar to those mentioned by teachers in the interview phase of this study, including the idea that some teachers did not see any reason to change, recalling Teacher A’s interview; and that there was insufficient communication between the test designers and the teachers, recalling the interview comments of Teacher C.

The study undertaken by Alderson and Hamp-Lyons, discussed above, led those researchers to the conclusion that:

The existence of a test by itself does not guarantee washback, either positive or negative — and ... Tests will have different amounts and types of washback on some teachers and learners than on other teachers and learners. (Alderson and Hamp-Lyons 1996:295–6)
In addition to this, this researcher would argue that evidence for washback may be found at some times and not at others; and, in some aspects of teaching and not in others, as the results of the observation phase indicate.

To state the obvious, evidence for washback may not always be found because it may not always be present. Where Alderson and Wall (1992:134) remarked on the lack of washback found during the observations, could this have been because some of the teachers being observed were like Teacher A and there was nothing to be observed? Or could they have been like Teacher D and the lesson observed like Lesson D1 and not D2? Where Wall (1996) remarked on the fact that the exam had so much impact on some aspects of teaching but so little on others (1996:348), could the teachers observed have been like Teacher B who had adapted the curriculum and assessment to meet his own needs? Washback may also manifest itself differently at different times or under different conditions, as shown in the difference between the two lessons of Teacher B and Teacher D. Thus different manifestations of washback may only be accessible through different methods of data collection, providing support for the use of triangulation in research of this kind (Bailey 1997:82).

Summary of findings
The findings of the survey phase of the research were largely inconclusive to the extent that they did not reveal evidence of washback. They provided, nevertheless, a picture of assessment practices in 1994, a useful (and necessary) starting point for the subsequent investigation. The interview phase was much more rewarding in its findings and uncovered a great deal of support for the idea that washback had occurred, although, importantly, it also uncovered a group of participants who claimed to have resisted making any changes as a result of the implementation. The data gathered through the interviews allowed for the refinement of the hypotheses to be investigated during the observation phase and began to lead the researcher towards the idea of an individual response to the implementation. The interview data also exposed the inappropriateness of the distinction being made between the implementation of the assessment and the curriculum.

The observation data provided the opportunity to examine various changes identified through the interviews in depth. The analysis of the lessons observed supported the hypothesis that washback had occurred in the teaching of Teachers B, C and D and that washback, as change, had been resisted by Teacher A, who showed no observable evidence of having adopted the curriculum or its assessment. The analysis of these data also showed differences between the teachers as individuals, each of those who had experienced washback showing evidence of this in different ways. These differences suggested that, at least in the case of classroom-based assessment, washback is a more idiosyncratic experience than had previously been proposed in the literature which, in turn, led to the models of teacher behaviour outlined at the conclusion of the interview phase. This, in turn, resulted in the proposal of a new model for washback, which appears below.

The collection of empirical evidence from classrooms means that it can confidently be claimed that, on the days they were observed, the behaviour of three of the four teachers did display evidence of washback — the effects of the introduction of the assessment on their teaching — to differing degrees and in differing manners. Each of these three teachers provided empirical evidence for the comments about change which they had made in their interviews, as well as the claims of other interview respondents. Teacher D showed evidence of changes that occurred only at certain times, particularly when she was focused on the assessment of the competencies, a change which she had described in her interview. She was labelled a ‘partial adopter’.

Teachers B and C showed a far more integrated approach to assessment, as well as demonstrating specific areas in the Content of their teaching which were related to the changes they had mentioned in their interviews. Teacher C had adopted a teaching Content based very largely on systemic-functional linguistics and genre theory, and had integrated the assessment into her teaching. Teacher C was labelled an ‘adopter’. In contrast, Teacher B had changed his teaching in specific areas, such as the adoption of genre-based teaching and the implementation and integration of the assessment throughout his teaching, while maintaining some of his previous practices. Teacher B was labelled an ‘adopter’.

Just as she had claimed in her interview, however, Teacher A demonstrated no observable evidence of a washback effect on her teaching. Not only did her teaching fail to show any evidence of washback but it also demonstrated those characteristics which some of the teachers interviewed had described as practices they had come to abandon, to a greater or lesser degree. It should be acknowledged, however, that this was neither laziness nor the result of a lack of awareness but rather a conscious decision on the part of that teacher, who had remained unconvinced of the benefits of the implementation of the assessment and who had, therefore, decided to resist. Teacher A described the assessment
A re-examination of the interview data allows for a grouping of the respondents according to their beliefs, assumptions and knowledge (‘BAK’: Woods 1996:195) and this, in turn, leads to the possibility of viewing the interview participants in terms of the models described above, based on their stated responses to the implementation of the CSWE and its assessment as an educational change. There were those who stated that they saw the implementation as positive or negative or both; those who described the change it represented as deep or superficial; and those who adopted the CSWE and its assessment immediately, others who adopted and adapted them as time passed and others who resisted the CSWE and its assessment as far as they were able within a mandatory system. From their comments, the interview respondents could be categorised as falling into three groups.

The first group were those whose educational and philosophical beliefs, assumptions and knowledge were in conflict with the theoretical and educational bases of the CSWE and its assessment. In terms of models of teacher response, these people were the resisters. This group often commented on the challenge to their expertise that the implementation represented, usually in terms of an underestimation of their skills and experience. This group also included those who most doubted the quality of the assessment and who stated that they only used the assessment to the extent they were required to do so.

The second group were those whose educational and philosophical beliefs were wholly or largely in accordance with, or even supported by, the theoretical and educational bases of the CSWE and its assessment. In terms of models of teacher response, these people were the adopters and, frequently, the early adopters. The members of this group had often taken part in pilot projects or had taken on leadership roles during the implementation, although this was not universally true. Most members of this group remarked that the change had not been profound for them personally, although some stated that the implementation had affected others deeply.

The third facet is that washback, as educational change, can be resisted, at least in the case of this type of assessment. The fact that the CSWE does not prescribe mandatory assessment tasks but gives only outcomes and examples of how these might be assessed adds an additional element of choice and, thus, a further possibility for resistance. The existence of Teacher A and her response to the implementation demonstrates that it is no longer sufficient to claim that ‘A test will influence teaching’ (ibid:120). This research has clearly demonstrated that (pace Alderson and Wall, ibid:121) while an assessment system may have a variety of washback effects for some teachers, it will not have the same washback effect for all teachers and, indeed, that it is possible to resist, consciously and successfully, the effects of the implementation of an assessment system which involved classroom-based, connected assessment.

Towards a new model for washback

Models of teacher behaviour

A re-examination of the interview data allows for a grouping of the respondents according to their beliefs, assumptions and knowledge (‘BAK’: Woods 1996:195) and this, in turn, leads to the possibility of viewing the interview participants in terms of the models described above, based on their stated responses to the implementation of the CSWE and its assessment as an educational change. There were those who stated that they saw the implementation as positive or negative or both; those who described the change it represented as deep or superficial; and those who adopted the CSWE and its assessment immediately, others who adopted and adapted them as time passed and others who resisted the CSWE and its assessment as far as they were able within a mandatory system. From their comments, the interview respondents could be categorised as falling into three groups.

The first group were those whose educational and philosophical beliefs, assumptions and knowledge were in conflict with the theoretical and educational bases of the CSWE and its assessment. In terms of models of teacher response, these people were the resisters. This group often commented on the challenge to their expertise that the implementation represented, usually in terms of an underestimation of their skills and experience. This group also included those who most doubted the quality of the assessment and who stated that they only used the assessment to the extent they were required to do so.

The second group were those whose educational and philosophical beliefs were wholly or largely in accordance with, or even supported by, the theoretical and educational bases of the CSWE and its assessment. In terms of models of teacher response, these people were the adopters and, frequently, the early adopters. The members of this group had often taken part in pilot projects or had taken on leadership roles during the implementation, although this was not universally true. Most members of this group remarked that the change had not been profound for them personally, although some stated that the implementation had affected others deeply.
The third group were those with more mixed feelings, since their beliefs, assumptions and knowledge were challenged by but not in complete conflict with the theoretical and educational bases of the CSWE and its assessment. In this group were found the partial adopters and the adaptors. These were also people for whom CSWE and its assessment offered some perceived benefit but who also had some doubts about the quality of the assessment or the quality of the implementation process. These doubts were not as strong as those of the resisters, however. This group can be divided into three subgroups.

In the first subgroup were those who found that they gained professional knowledge from the implementation, for example in the assessment field, or that the CSWE and its assessment gave them guidance, ‘focus’ or an increased authority in the classroom. The members of this group were adopters, although not necessarily early adopters. The second subgroup were those who found an unexpected, though not necessarily personal, benefit in the implementation. This included those who spoke about how the CSWE and its assessment had led other teachers to teach literacy instead of ignoring it, or to use more ‘acceptable’ teaching methods. The members of this group were mainly adopters. Those in the third subgroup were the adaptors, teachers who found that they were able to adapt the CSWE and its assessment to meet their needs and to suit their teaching styles. This group also included those who stated that their attitude to the CSWE and its assessment had become more positive with the passage of time. There was considerable overlap between the second and third subgroups.

These models of teacher behaviour would appear to be related to the concept of the belief system, characterised by Linde as a ‘... set of beliefs which is coherent, which is focused around some central issue, and which is not held by everyone in a given culture’ (Linde 1980, in Woods 1996:69). As Woods (ibid) writes:

Teachers ‘interpret’ a teaching situation in the light of their beliefs about the learning and teaching of what they consider a second language to consist of; the result of this interpretation is what the teacher plans for and attempts to create in the classroom.

Reconceptualising washback

These findings led to the proposal of a new model for washback (after Woods 1996:51), informed both by previous views and latest developments in washback research. This proposal explicates washback as a construct, taking into account teachers’ belief systems and consequent responses to change. Three models of washback appear below: the traditional model, which pre-dates Alderson and Wall (1993); a second model, which relates to current writing about washback (eg Shohamy et al 1996); and a proposed third model, which relates washback to curriculum innovation and teachers’ beliefs, assumptions and knowledge.

Figure 4.1: Traditional washback theory: A stimulus-response model

The traditional washback view pre-dated Alderson and Wall (1993). This view held that the introduction of any test would necessarily lead to a washback effect, a single and uniform response, which might be negative or positive (for example Vernon 1956; Wilkinson 1968; Fransson 1984; Crooks 1988; Paris et al 1991; Smith 1991). Where this was amplified in the literature, the type of response depended on qualities of the test rather than on the teachers involved: hence the consistency of the teachers’ response. From this view came the concept of washback validity (Morrow 1986) and working for washback (Swain 1985; Hughes 1989). Observational and other evidence-based washback research were not characteristic of this model of washback.

Figure 4.2: 1990s view of washback: A ‘black box’ model

Unlike the previous model, his model has been informed by objective data gathering, with observers gathering empirical evidence which discounts the first view. Following the publication of Alderson and Wall’s (1993) article, a series of evidence-based, observational washback studies were undertaken. In these studies, which have been cited above and include Blewchamp (1994), Alderson and Hamp-Lyons (1996), and Watanabe (1996), researchers collected empirical evidence for washback, primarily through classroom observation. The data collected indicated the existence of individual responses to the implementation of a variety of tests, an idea taken further by Wall (1996), who proposed a series of reasons for these responses which were related to the concept of teachers as individuals in a context of educational change. The revelation that the teachers
involved in these studies did not all respond in the same way discounted the first model, since they indicated that a single response to the introduction of a new test was not inevitable.

```
        New test
          ↑         teachers         models of responses
            BAK
```

**Figure 4.3: Proposed view of washback: A curriculum innovation model**

Like the second model, this model has also been informed by objective data gathering, with observers gathering empirical evidence which discounts the first view. This view has been further informed by analysis of the data for patterns of behaviour, following models derived from curriculum innovation. This model embodies the view that a qualitative analysis of teachers’ responses to the introduction of a new test or assessment system may reveal patterns in their responses and that these patterns may be similar to Markee’s (1997) models of response to educational change and the models of teacher assessment proposed by McCallum et al (1995). The issue here may best be summarised by the following quotation:

... a solely behavioral model is conceptually incomplete. It cannot account for predictable variations in teachers' behavior arising from differences in their goals, judgments, and decisions. (Shavelson and Stern 1981:455)

This comment, which in the original related to teachers’ pedagogical intentions and behaviours, relates equally well to washback. Unless it were demonstrated otherwise through empirical evidence, it does not seem an acceptable proposition that washback should be different from all other types of educational change. Indeed, recent research discussed above supports the view being proposed here that teachers will not demonstrate simplistic stimulus-response behaviours but will rather respond to the stimulus provided by the implementation of a new test or assessment system as they would to any other educational change, their responses being governed by their beliefs, assumptions and knowledge. For this reason, the early model of washback described above should be discounted, even in high-stakes testing. The second model, essentially a summary of Blewchamp (1994), Shohamy et al (1996), Alderson and Hamp-Lyons (1996) and Watanabe (1996), takes as its basis the individual responses these studies have noted as a result of observation. This view might perhaps best be summarised by Blewchamp’s comment (1994:47), already quoted above, that his ‘... data also suggest that the differences between teachers are as important a factor as differences between (the types of classes he examined) ... in determining what happens in the classroom’.

The third model shows the results of analysis of individual responses for patterns in or models of teacher behaviour, as has been proposed in this research. This model is grounded in the notion that washback is a form of educational change, as Wall (1996) argued, and from this that models proposed for other areas of educational change (for example Markee 1997; McCallum et al 1995) might be applicable here, since studies of washback and curriculum innovation have in common the examination of the impact of educational change on teaching.

Studies of curriculum innovation have long explored the concept of an individual response to change and have consequently proposed models of teacher behaviour. Washback research, on the other hand, has traditionally been more concerned with the examination of large-scale responses to change. In recent years, washback research has begun to apply a rigour to its investigations, while an intelligent application of complementary qualitative and quantitative methods has led to rapid and exciting developments in refining and redefining washback. Such studies have demonstrated that washback is a more complex phenomenon than it had previously been thought to be, a concept which included the notion that the teacher was an important factor in the occurrence of any washback effect.

This study has demonstrated the importance of the teacher as a medium through which any change brought about by the implementation of assessment must pass. It has found evidence of washback in a connected assessment system, as well as evidence that the effect of the implementation of assessment may be resisted. This study has also proposed that teachers appear to respond in ways which may be described in terms of models or patterns of behaviour derived from curriculum innovation theories. It has further proposed that teachers’ responses are related to their attitudes towards and experiences of the implementation of the assessment of the CSWE; their perceptions of the quality of the assessment; the extent to which the assessment represented a departure from their previous practices; and their attitudes to change itself.
These findings have implications for future washback research, as an area which may be investigated further in other contexts, and for previous washback research, as a theory which might explain some of its findings. It has implications for those who wish to maintain the notion of washback validity, since it raises the question of whether this type of validity can be controlled when the effects of an implementation can be resisted. It also has implications for those who wish to ‘work for washback’, since it indicates to those who wish to implement new assessment systems or tests that there are lessons to be learnt from curriculum innovation which should be taken into account in an attempt to ensure the successful implementation of innovations in the field of testing and assessment.

Endnotes
1 Not all agree with this view (Mathison 1988, in Merriam 1988).
2 South Australia later suspended use of CSWE and was omitted from further phases of the study.
4 cf Chapter 6 of Spada and Fröhlich (1995) which contains 11 examples of its use in a variety of language teaching studies.
5 Teacher B also employed this technique in teaching the casual conversation.
6 Where a teacher used both a minimal text and an extended text simultaneously (for example in listening to a casual conversation and finding vocabulary items) the text was coded as minimal + extended.

References
Adult Literacy Information Office 1996. Professional development kit for the national reporting system. Foundation Studies Training Division, Western Sydney Institute of TAFE: National Staff Development Committee
Alcorso C and M Kalantzis 1985. ‘Consultancy E: The learning process and being a learner in the AM EP.’ Paper prepared for the Committee of Review of the Adult Migrant Education Program. Wollongong: Centre for Multicultural Studies
Alderson, J C and D Wall 1992. The Sri Lankan O-level English language evaluation project: Fourth and final Report. Lancaster: Department of Linguistics and Modern English Language Institute for English Language Education
Bailey, K M 1997. ‘Washback in language testing.’ Unpublished manuscript, Graduate School of Languages and Educational Linguistics, Monterey Institute of International Studies
Blewchamp, P 1994. ‘Washback in TOEFL classrooms: An exploratory investigation into the influence of the TOEFL test on teaching content and methodology.’ Unpublished dissertation, University of Lancaster, Department of Linguistics and Modern Languages


Chappell, C 1996. ‘Quality and competency standards’. Prospect, 11, 1: 59–70


Ingram, D E and E Wylie 1984. Australian second language proficiency ratings. Canberra: Department of Immigration and Ethnic Affairs


Joyce, H 1993. ‘Competencies in focus’. Interchange, 22: 1–7


National Training Board 1993. ‘UK update — access to assessment’. National Training Board Network, 12: 24


Raethel, S 1996. ‘20% of teachers suffer stress disorders: study’. Sydney Morning Herald, July 1: 2


The impact of assessment in the Certificate in Spoken and Written English

Studies in immigrant English language assessment
Acknowledgments

My thanks first to my supervisor, Geoff Brindley, for his endless patience and wise counsel and for putting his vast knowledge of the field at my disposal. Second, to all the participants in this study, especially Teachers A, B, C and D, without whom I could not have undertaken this research. Third, I’d like to thank my colleagues at NSW AMES, especially Joyce Murray and Bruce Hamilton. I would also like to acknowledge the staff of the N CELT R Resource Centre at Macquarie University, to whose assistance my reference list attests. My thanks also to my friends John Pickett and Ross Donald for their encouragement.

Finally, my thanks go to my family and to Sue, who now knows more about washback than she might care to, and without whose support this study would never have been completed.
Question One: I want to ask you about the extent to which you use competency-based assessment of the CSWE in your classroom. Could you please describe your use of competency-based assessment of CSWE in your classroom? When you describe it, please think about the following questions:

1a) How often do you use it?

1b) In what ways do you use it?

Question Two: Do you think the implementation of the assessment of CSWE has been positive or negative or both? In what ways?

Question Three: Change has sometimes been described like this: ‘change can be very deep, striking at the core of learned skills and beliefs and conceptions of education, and creating doubts about purposes, sense of competence, and self-concept’ (Fullan 1991:45).

To what extent do you think this describes the implementation of the assessment of CSWE? Why?

Question Four: Now I am going to ask you about some of the ways that some teachers said their teaching had changed since the implementation of the assessment of CSWE.

4a) In what ways have the structure and organisation of your classroom changed since the implementation of assessment of CSWE?

4b) Has the implementation of assessment of CSWE led to your using any different teaching methods?

4c) Has the implementation of assessment of CSWE led to your using any different teaching materials?

4d) Have you observed any other changes in your own behaviour as a teacher since the implementation of assessment of CSWE?

4e) Have you observed any changes in your students’ behaviour since the implementation of assessment of CSWE?

4f) Have you observed any (other) changes in your classroom which you feel have been brought about by the implementation of the assessment of the CSWE?

Question Five: Thinking back on the answers you have just given, do you think these changes are the direct result of the implementation of the assessment of CSWE?

Question Six: Is there anything you would like to add?