Exploring task difficulty in ESL listening assessment
Geoff Brindley and Helen Slatyer
Language Testing 2002 19: 369
DOI: 10.1191/0265532202lt236oa

The online version of this article can be found at:
http://ltj.sagepub.com/content/19/4/369

Published by:
SAGE
http://www.sagepublications.com

Additional services and information for Language Testing can be found at:

Email Alerts: http://ltj.sagepub.com/cgi/alerts
Subscriptions: http://ltj.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav
Citations: http://ltj.sagepub.com/content/19/4/369.refs.html

>> Version of Record - Oct 1, 2002
What is This?
Exploring task difficulty in ESL listening assessment

Geoff Brindley and Helen Slatyer Macquarie University, Sydney, Australia

This article reports on an exploratory study that investigated the comparability of listening assessment tasks used to assess and report learning outcomes of adult ESL learners in Australia. The study focused on the effects of task characteristics and task conditions on learners’ performance in competency-based listening assessment tasks that require learners to demonstrate specific listening behaviours. Key variables investigated included the nature of the input and the response mode. Quantitative and qualitative analyses of test scores suggest that speech rate and item format influence task and item difficulty. However, the complexity of the interaction between text, item and response makes it difficult to isolate the effects of specific variables. Implications of these findings for assessment task validity and reliability are considered and practical consequences for assessment task design in outcomes-based systems are discussed.

I Introduction

1 Outcomes-based approaches to assessment and reporting

In recent years, language learning programmes worldwide have come under increasing pressure to provide more explicit information on educational outcomes as part of their accountability requirements. One of the consequences of this trend has been the widespread adoption of descriptors of achievement known as ‘standards’, ‘benchmarks’, ‘competencies’ or ‘attainment targets’ as a basis for assessing learner progress and reporting learning outcomes to external stakeholders (Brindley, 1998a; 2001; McKay, 2000). Although outcomes-based approaches offer a number of advantages, including closer alignment of curriculum and instruction and potentially better communication between stakeholders, a range of concerns have been expressed in relation to the validity and reliability of the assessments that are used to determine levels of achievement (Brindley, 1998a; 2001; McKay, 2000; 2001). One particularly thorny
problem concerns the comparability of teacher-developed assessments: if different teachers are responsible for selecting or creating assessment tasks, then how can we ensure that scores on tasks that are assessing the same outcome have the same meaning? Moreover, given the well-documented influence of variations in task characteristics and task conditions on language test performance (Bachman and Palmer, 1996; Skehan, 1998), to what extent is it possible to develop parallel tasks that elicit the same range of language features under the same conditions of administration? In outcomes-based systems, lack of comparability across tasks is a potentially serious threat to validity and reliability that could lead to invalid inferences concerning learners’ achievement of the targeted outcomes.

This article reports on an exploratory study that set out to investigate these questions in the context of the Certificates in Spoken and Written English (CSWE), an outcomes-based curriculum and assessment framework used within the federal government-funded Adult Migrant English Program (AMEP) in Australia (New South Wales Adult Migrant English Service; NSW AMES, 1998). In an attempt to throw some light on the question of task comparability, the study focused on the manner in which variations in task conditions and task characteristics influence the difficulty of listening competency assessment tasks. First, the CSWE framework is described and the rationale for the present study is presented. Some of the particular challenges that face task designers in competency-based assessment systems are then briefly outlined. Following this, the specific variables that were investigated are discussed and the methodology for the study is described. The results of quantitative and qualitative analysis of item difficulties and of the assessment tasks used are then discussed. Finally, practical implications for assessment task design are explored.

2 Background: The Certificates in Spoken and Written English

The CSWE framework encompasses a range of language skills that are necessary for adult immigrants and refugees in Australia to be functionally proficient in English in the community, at work and in further study. It describes learning outcomes at four different language proficiency levels in terms of language competencies. The proficiency levels are roughly equated to the following range of levels on the Australian Second Language Proficiency Ratings (ASLPR),1

---

1The ASLPR has been revised in recent years and is now known as the International Second Language Proficiency Ratings (ISLPR) (Ingram and Wylie, 1997). The ASLPR 0 to 2 range spans approximately the Novice Low to Intermediate Mid levels on the ACTFL scale (Breiner-Sanders et al., 2000).
a rating scale that is used in conjunction with an oral interview to assess learners’ proficiency at entry to the AMEP (Ingram and Wylie, 1984).

<table>
<thead>
<tr>
<th>CSWE I</th>
<th>beginner</th>
<th>ASLPR 0 to 0+</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSWE II</td>
<td>post-beginner</td>
<td>ASLPR 1– to 1</td>
</tr>
<tr>
<td>CSWE III</td>
<td>intermediate</td>
<td>ASLPR 1+ to 2</td>
</tr>
<tr>
<td>CSWE IV</td>
<td>advanced</td>
<td>ASLPR 2 and above</td>
</tr>
</tbody>
</table>

Each Certificate consists of a number of language competencies in oral interaction, reading and writing, typically expressed as ‘can-do’ statements (can participate in a casual conversation; can respond to spoken instructions; can read an information text; can write a report, etc). The competencies are broken down into:

- elements, which describe the skills and knowledge involved in the performance;
- performance criteria, which set out the obligatory elements of performance against which the learner is to be assessed;
- range statements, which specify the conditions under which the performance occurs (e.g., the time available for the task, the amount of support and resources available, etc).

In addition, an evidence guide provides sample assessment tasks. Table 1 shows the description of the listening competency that is the subject of the present investigation. Teachers decide which competencies they will teach and assess according to learners’ ability and needs. When learners are considered ready, they are assessed in that competency using assessment tasks that aim to elicit evidence of their ability to use or process language under the conditions described in the range statements. In order to qualify for the award of a competency, learners must demonstrate mastery of all of the performance criteria for the competency in question. Assessments are carried out by teachers in the classroom, using tasks that they have developed themselves, or with the aid of exemplar tasks that accompany the CSWE. A set of assessment support materials has also been developed to assist teachers in the design and evaluation of assessment tasks (Christie and Delaruelle, 1997). The award of a certificate is contingent on the learners’ attainment of a given number of competencies in each language skill, the total number of which may range from 13 to 16, depending on the level of the certificate.

At the end of a course of instruction, teachers are required to report on learners’ achievement of CSWE competencies to the funding authority, the Commonwealth Department of Immigration and Multicultural Affairs (DIMA). Reporting is carried out via a national data management system, ARMS (AMEP Reporting and Management
Table 1  Competency description: Certificate III Competency 3 (Competency 3: Can demonstrate understanding of an oral presentation report)

<table>
<thead>
<tr>
<th>Elements</th>
<th>Performance criteria</th>
<th>Range statements</th>
<th>Evidence guide</th>
</tr>
</thead>
<tbody>
<tr>
<td>i. can identify main ideas</td>
<td>• identifies main ideas</td>
<td>• familiar and relevant topic</td>
<td>Sample tasks: Learners listen to and answer questions on a short lecture/classroom presentation; e.g.:</td>
</tr>
<tr>
<td>ii. can identify explicitly stated information</td>
<td>• identifies specific information which supports or elaborates on main ideas</td>
<td>• text segment for assessment approximately 2 minutes in length</td>
<td>• speaker from local employment agency</td>
</tr>
<tr>
<td>iii. can identify supporting arguments</td>
<td></td>
<td></td>
<td>• speaker from jobtraining program</td>
</tr>
<tr>
<td>iv. can identify implicit ideas/information</td>
<td></td>
<td></td>
<td>• talk at education centre</td>
</tr>
<tr>
<td>v. can identify logical relationships</td>
<td></td>
<td></td>
<td>• speaker from union/professional organization</td>
</tr>
<tr>
<td>vi. can demonstrate understanding of vocabulary</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: NSW AMES, 1998

System), which enables DIMA to monitor aggregate CSWE outcomes in AMEP courses across Australia.

Certificate attainment is a key element in the AMEP’s public accountability requirements since average competency gains are used by DIMA as one of the key outcome indicators for reporting to the Federal Department of Finance on the performance of the AMEP (Ross, 2000). Thus, the inferences made on the basis of CSWE assessments play an important role in funding decisions affecting the Program’s future.

II Assessing achievement in a competency-based system: issues and challenges

1 Achieving consistency

Since certificate attainment in the CSWE is largely determined on the basis of teacher-developed assessment tasks, the quality and replicability of these tasks are important factors affecting the dependability of
the information on learner outcomes. Tasks that are aimed at assessing the same competency thus need to be designed so that they elicit the same behaviours and are administered in the same way. To do this, they should be accompanied by the same degree and type of contextual support and present a comparable level of cognitive challenge for learners. Secondly, there needs to be a high level of consistency in the way that assessors classify competencies as achieved or not achieved. This requires a common interpretation and application of the performance criteria within each competency.

2 Identifying and controlling variability

Evidence from studies of CSWE assessments to date, however, suggests that achieving consistency may be a difficult goal to attain under real-life constraints. In this context, considerable variability has been found in the way in which the assessment tasks are designed, administered and rated (Brindley, 2000; 2001; Smith, 2000; Wigglesworth, 2000a). A common finding that emerges from all of these studies is the extent to which variations in task characteristics and task conditions may influence learners’ linguistic output. For example, Wigglesworth (2000a) found that interlocutor behaviour in oral assessment tasks influenced both the quality and quantity of the language produced by learners and, hence, the scores that they obtained. CSWE writing tasks assessing the same competency were also found to be of differential difficulty when subjected to many-facet Rasch analysis (Brindley, 2000). In order to design valid competency assessment tasks that remain stable across different administrations, therefore, task effects need to be understood, and, as far as practicable, controlled.

III The study

1 Aims

The research studies described above have provided useful information on factors affecting task variability that has informed the design of speaking and writing tasks (Wigglesworth, 2000b). However, it remains unclear to what extent receptive competencies are subject to the same degree of variability. It was therefore decided to mount a parallel study focusing on listening competencies.

The principal aims of the present study were:

- to identify key task characteristics and task conditions that were most likely to affect the difficulty of CSWE listening assessment tasks;
to determine the extent to which these characteristics and conditions were accounted for in the range statements;

• to investigate the impact on test scores of systematically varying task characteristics and task conditions and, in cases where clear effects were noted, to explore possible reasons for differences in task and item difficulty.

2 The listening competency

The competency that was the focus of this study was Competency 3 in Certificate III: ‘Can demonstrate understanding of an oral presentation/report’ (see Table 1). The ‘Range statement’ describes the specific characteristics of the task and the conditions under which the task is to be administered.

This is the only non-interactive listening competency at Certificate III level. Other oral interaction competencies involving listening skills at this level are:

• Competency 4: Can respond to complex spoken instructions;
• Competency 5: Can negotiate a complex/problematic spoken exchange;
• Competency 6: Can participate in a casual conversation.

Learners may preview questions before they hear the text. No indication of the time recommended for preview is given, nor whether learners may use their dictionaries during this time to look up vocabulary items. No specific item types are recommended, but multiple-choice items are specifically excluded.

3 Selection of variables

The variables that were identified for initial investigation were chosen for both practical and theoretical reasons. First, from the practical standpoint of comparability, it was important to establish whether there were particular task conditions or task characteristics whose omission from the range statements might introduce construct-irrelevant variance into test scores. If this turned out to be the case, then steps could be taken to incorporate these factors into subsequent revisions of the competency specifications. At the same time, from a theoretical perspective, the study presented an opportunity to investigate some of the hypotheses that have been advanced in the research literature concerning those variables that affect second language listening comprehension.
4 Investigating listening task difficulty

a Terminology: For the purposes of this discussion we adopt the definition of ‘task’ proposed by the Association of Language Testers in Europe (2001):

- a combination of rubric, item and response. For example, a reading text with several multiple-choice items, all of which can be responded to by referring to a single rubric.

As Davies et al. (1999:196) point out, the terms ‘item’ and ‘task’ tend to overlap:

Where a distinction is made between them, there is usually an implication that an item is smaller and less complex, while a task is larger and more complex. A test would typically include either a large number of relatively short items or a small number of relatively complex tasks.

A task for assessing the listening competency described above would thus consist of the input text that learners hear, along with the accompanying set of items and instructions.

b Factors affecting task difficulty: A large number of factors that may affect listening task difficulty have been identified by researchers (Buck, 1990; 2001; Rost, 1990; Henning, 1991; Dunkel et al., 1993; Flowerdew, 1994; Rubin, 1994; Freedle and Kostin, 1996; Nissan et al., 1996; Jensen et al., 1997; Brindley, 1998b; Bejar et al., 2000). Amongst these are:

- the nature of the input: speech rate, length of passage, syntactic complexity, vocabulary, discourse structure, noise level, accent, register, propositional density, amount of redundancy, etc.;
- the nature of the assessment task: amount of context provided, clarity of instructions, response format, availability of question preview, etc.; and
- the individual listener factors: memory, interest, background knowledge, motivation, etc.

In addition, a small number of research studies have focused more specifically on particular item characteristic variables that may impact on item difficulty (Freedle and Kostin, 1996; Nissan et al., 1996; Jensen et al., 1997; Buck and Tatsuoka, 1998). Features that have been found to affect the difficulty of listening comprehension items across a range of task types include:

- amount of lexical overlap between the text and the response format (Freedle and Kostin, 1996; Buck and Tatsuoka, 1998);
- length of text preceding the information required to respond (Jensen et al., 1997);
length of required response (Jensen et al., 1996; Buck and Tatsuoka, 1998);
repetition of tested information (Freedle and Kostin, 1996; Jensen et al., 1997; Buck and Tatsuoka, 1998);
whether responses and repetitions of information are verbatim or paraphrases (Jensen et al., 1997).

c Key variables investigated: Given the practical difficulties involved in investigating the effects of all of these variables simultaneously, it was decided to narrow the range of investigation to five key variables that seemed most relevant in the context of the CSWE listening competency under investigation and to try to isolate their specific effects on test scores. Of these variables, four were related to the nature of the input and one to the response mode. They were as follows:

• Speech rate: Speech rate has been shown by numerous research studies to have a major influence on second language listening comprehension (e.g., Griffiths, 1990; 1992; Zhao, 1997). Buck (1990: 91–93) reviews a range of studies, all of which suggest that faster rates of delivery can reduce comprehension. The CSWE competency specifications, however, make no mention of speed of delivery, leaving open the possibility that learners who are exposed to faster rates of delivery would be disadvantaged.

• Text type: There is some evidence to suggest that the degree of ‘orality’ (that is, the extent to which the text includes features of spoken as opposed to written language) affects listening test scores. Shohamy and Inbar (1991), for example, found that listening test passages ranged in difficulty along a continuum that corresponded to the degree of ‘orality’ of the text in question, with the more ‘oral’ text types such as consultative dialogues being easier. Other studies have also suggested that dialogues containing negotiated discourse and recycling of propositions can aid comprehension (Rost and Ross, 1991; Ross and Langille, 1997). We therefore decided to investigate whether presenting the same information in the form of a dialogue resulted in any variation in test scores.

• Number of hearings: Many public tests allow for a text to be heard only once. Some, however, allow for two hearings at lower levels (e.g., some of the Cambridge examinations). Skehan (1998) suggests that one hearing only would increase the cognitive load by making greater demands for online processing, thereby making a task more difficult. CSWE Competency 3 specifies that only one hearing should be allowed. However, given the proficiency level
of the learners in question (lower to mid-intermediate) and concerns on the part of some teachers that one hearing may represent an overly demanding processing load, we wanted to investigate the effects of allowing an extra hearing of the text.

- Input source (live vs audio-recorded): As noted above, the competency specifications treat ‘live’ (i.e., text spoken by the teacher or an invited presenter), video-recorded and audio-recorded sources as interchangeable. However, these modes of presentation are potentially quite variable; for example in terms of contextualization, discourse structure and propositional density and thus might affect test scores.

- Item format: Some researchers have found that different item formats may make differing processing demands on candidates’ performance in listening tests (Berne, 1993; Hansen and Jensen, 1994; Nissan et al., 1996). A study by Berne (1993) showed that subjects performed better on multiple-choice questions than on either an open-ended or cloze task, suggesting that items requiring only recognition are easier than those requiring retrieval and production. Most of the research to date, however, has been on multiple-choice items, and the effects of other commonly used formats such as short answer questions or information transfer tasks remain largely unexplored. The CSWE does not specify any particular item formats (apart from proscribing multiple choice), and it was therefore decided to examine whether varying the item format resulted in any differences in test performance.

5 The assessment tasks

Three assessment tasks were used. All were based on tasks that had originally been developed by AMEP teachers:

1) Task 1, ‘Education’, which was administered to all candidates as a control task. This consisted of a recorded monologue of approximately two minutes in length concerning the Australian education system, accompanied by 10 sentence completion items requiring the candidates to write a few words in English.

2) Two manipulated tasks. A baseline version of each of these was developed conforming to the CSWE specifications as follows:
   a) Task 2, ‘Jobs’: This consisted of a recorded monologue at normal speed – operationalized here as 180 words per minute (wpm) – concerning different ways of finding a job. The response format used was the same as for the ‘Education’ task. One hearing was allowed.
   b) Task 3, ‘Dogs’: This consisted of a recorded passage at normal speed (180 wpm) in the form of a monologue about
the Guide Dog Association in Australia, its history, and the work it does. The response format was the same as for the other two tasks. One hearing was allowed.

Alternative versions of each of these two tasks were then developed in which one task characteristic or task condition was varied. These changes are summarized in Table 2.

- **Item format:** Task 2B used the baseline version of the ‘jobs’ text. The response format was a table that learners were required to fill in while listening. The table was designed to reflect the rhetorical structure of the text (see Figure 3 in Section IV). The column headings reflected the key information in the text (different ways of finding a job) and the rows (advantages and disadvantages of each of these) were signalled by discourse markers in the text (e.g., ‘one good thing about . . . ’; ‘but one problem is . . . ’). Each item required a few words for the response (generally a maximum of three words); Task 3B used the baseline version of the ‘dogs’ text, but the items were in the form of short answer questions (SAQs). The SAQs were a reformulation of the sentence stems in the sentence completion formats.

- **Text type:** Task 3C was a ‘dialogue’ version of the baseline ‘dogs’ text with the baseline item format (sentence completion). The dialogue was designed to include more redundancy and more fillers.

**Table 2** Assessment tasks and variations

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Variations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Task 1: (Control) Education</strong> (Baseline version)</td>
<td>Baseline version of the text, baseline version of items</td>
</tr>
<tr>
<td><strong>Task 2: Jobs</strong> A: Baseline version</td>
<td>Baseline version of the text, baseline version of items</td>
</tr>
<tr>
<td>B: Item format: table</td>
<td>Baseline version of the text, items in table format</td>
</tr>
<tr>
<td>C: No. of hearings: 2</td>
<td>Baseline version of the text, baseline version of items; 2 hearings</td>
</tr>
<tr>
<td>D: Vehicle: live</td>
<td>Baseline version of the text, baseline version of items; live version</td>
</tr>
<tr>
<td><strong>Task 3: Dogs</strong> A: Baseline version</td>
<td>Baseline version of the text, baseline version of items</td>
</tr>
<tr>
<td>B: Item format: SAQ</td>
<td>Baseline version of the text, SAQ items</td>
</tr>
<tr>
<td>C: Text type: dialogue</td>
<td>Dialogue version of the text, baseline version of items</td>
</tr>
<tr>
<td>D: Speech rate: faster</td>
<td>Baseline version of the text, faster speech rate, baseline version of items</td>
</tr>
</tbody>
</table>
This version was also longer than the baseline text (764 compared to 604 words).

- Task conditions: all other versions (Tasks 2C, 2D and 3D) used the baseline version of both text and items, but varied the task conditions (two hearings, live, and faster speech rate [200+ wpm]).

6 Procedure

Various combinations of the three tasks were administered to 284 adult ESL learners enrolled in Certificate III in teaching centres across three states in Australia (New South Wales, Victoria and Queensland). The control task was undertaken first, followed by one version of the ‘jobs’ task and one version of the ‘dogs’ task. The latter two were randomly assigned in an attempt to control for order effects. Time was allocated for question preview and dictionary use was permitted to look up vocabulary during the preview time. All listening passages were audio-recorded, except the live version which was read from a script by the teacher. Two minutes at the end of the task were allotted for learners to check their work.

The test responses were marked by a team of eight experienced ESL teachers. Answer keys were tested on a range of samples and decisions were collectively made about the acceptability of responses. During the marking process, any responses that did not correspond to the answer key were submitted to the group of markers and a collective decision was reached on the acceptability of the response. Dichotomous scoring was used, with each response being accorded correct or incorrect status. Missing or incomplete items were marked as incorrect.

IV Results

1 Analysis of combined test scores

As an initial step, an analysis of the scores was carried out using the Rasch-based Quest program (Adams and Khoo, 1993) to obtain person ability and item difficulty estimates. For purposes of the analysis, all of the test forms were combined and treated as a single test containing 89 items, with the control task providing a linking set of common items. This enabled the scores from all versions of all tasks to be placed on the same ability/difficulty scale. Mean item difficulty estimates for each version of the tasks were then used as an initial broad indicator of differences in task difficulty.

2One item was deleted from Task 3B after piloting.
The item-ability map showed that the tasks were well matched to the range of candidate ability. Item difficulties ranged from to –2.45 to 2.9 logits, with a mean of zero and a standard deviation of 0.93. Using the criteria that are built into the Quest program (infit mean square values outside the range of roughly 0.75 to 1.3), no misfitting items were identified, although four items were signalled as overfits, indicating little variability in responses. Person abilities ranged from –3.23 to 2.99 logits, with a mean of –0.10 and a standard deviation of 1.15. The Rasch reliability of case estimate (equivalent of KR20) was .86 for the combined tasks. Mean item difficulty estimates for each version of the listening tasks are given in Table 3.

2 Effect of variables on overall task difficulty

Only two of the variables appeared to make the tasks notably easier or more difficult. It can be seen that the ‘jobs’ task is easiest when the ‘table’ format is used and the ‘faster’ version of the ‘dogs’ task was the most difficult. Using the logit to probability conversion table provided by Wright and Linacre (1991), this suggests that:

- If the ‘table’ item format were used rather than the sentence completion, a candidate’s chances of achieving the competency would be increased by approximately 30%; and
- If a borderline candidate were administered the faster version of the ‘dogs’ passage, he or she would have about a 30% less chance of being awarded the competency than if he or she were given the version at ‘normal’ speed.
Figure 1  Combined item difficulty estimates: ‘jobs’ task

3 Variations in item difficulty

In order to examine the behaviour of individual items within each version of the two tasks, item difficulty estimates for all versions of each task were plotted onto two graphs (Figures 1 and 2). Looking at the variations in item difficulty across different versions of the two tasks, it can be seen that the mean item difficulty estimates mask a good deal of variability in the difficulty of the items within each of the versions. For example, even though the item difficulty estimates suggest that the ‘table’ version of the ‘jobs’ task was the easiest overall, some items (4, 5 and 6) were more difficult in that version than they were in other versions (see Figure 1). Similarly, although the ‘faster’ version of the ‘dogs’ task appeared to be the most difficult, Items 4 and 6 were easier in this version than in some of the other versions (see Figure 2).

Figure 2  Combined item difficulty estimates: ‘dogs’ task
4 Analysis of individual items

It was apparent that the broader input and response variables that we had tried to manipulate would not explain why particular items were more or less difficult in the different versions of the passages. As Buck and Tatsuoka (1998: 125) noted in their investigation of listening test item attributes:

The empirical attributes are far more rooted in the actual characteristics of the individual items themselves – what we might call a ‘nuts and bolts’ level of attribute definition – whereas the theoretical-speculative attributes are much more concerned with complex higher-order categories, or linguistic generalities.

We therefore decided to investigate the characteristics of individual items more closely in an attempt to isolate those specific factors at the item level that may be affecting learner performance. To this end, all the items in each version of the manipulated tasks (59 items overall: 30 items for the ‘jobs’ task and 29 for the ‘dogs’ task) were examined.

The analysis of the items focused on three structural components of listening assessment tasks that have been identified in the research literature as influencing item difficulty:

- The necessary information (NI): this is defined by Buck and Tatsuoka (1998: 134) as ‘the information in the text which the listener must understand to be certain of the correct answer’. In the case of the tasks considered here, which involve extended listening passages and written responses in the second language, the NI consists of the information needed to complete the sentence or table, that is, the words contained in the response, either verbatim or in paraphrase. The location of the NI and its linguistic characteristics were found by Buck and Tatsuoka (1998) to be key factors affecting item difficulty and candidate responses.
- The surrounding text: this refers to the text immediately surrounding the necessary information, from the point where the listener is first alerted to the upcoming information by a verbal cue up to the response or final repeat of the response if repetition is involved. The characteristics of this part of the text are identified by Buck and Tatsuoka (1998: 134) as having a greater effect on item difficulty than the characteristics of the whole text (see also Buck, 1991a).
- The stem: this is defined for the purposes of this study as the written text on the answer sheet which test-takers have in front of them as they listen and which serves both as a listening guide and a structure for the written response (Buck, 1991a). In listening comprehension tasks such as the ‘jobs’ and ‘dogs’ tasks, the stem
would be the beginning of the sentence to be completed (baseline versions), the short answer question (SAQ) to be answered or, in the table version, the structure of the table that is used to write the answer.

As an illustration of the interaction between these components and the effects of this interaction on item difficulty, an analysis of the easiest and most difficult items (Item 6 in the ‘dogs’ task and Item 10 in the ‘jobs’ task respectively) is presented below.

a Easiest item: Item 6 in the ‘dogs’ task was the easiest item for all versions of both tasks, with logit values ranging from 1.62 to –2.43. In this item, the necessary information and surrounding text are in the slowest part of the text (157 wpm, corresponding to Griffiths’ [1990] ‘slow’ rate) both in the baseline version and faster version [200 wpm]). The stem is constraining, allowing only a noun or noun phrase and has a simple SV grammatical structure in the baseline version. The SAQ version, which contained an open ended (wh-) question, was slightly more difficult.

- Stem and response:
  - Baseline version: In 1962 the Australian headquarters moved to . . . (Melbourne)
  - SAQ version: Where did the Australian headquarters of the Australian Guide Dog Association move to? (Melbourne)

The correct response consisted of one word, which was both familiar (the name of the second largest city in Australia) and salient in the prosody (strong word stress on a proper noun).

- Surrounding text:
  - Lexical overlap with the text: In 1962 the headquarters for the Blind Dog Association moved from Perth to . . . (Melbourne)
  - Repeat: and now blind people attend the National Guide Dog Association Training Centre in Melbourne.

The stem was strongly cued in the text with a high degree of lexical overlap. There was also a verbatim repetition of the response, a feature that was found by Jensen et al. (1997) to be a predictor of item difficulty.

b Most difficult item: In contrast, item 10 in the ‘jobs’ task was the most difficult item in all versions except for the ‘table’ format version where it was easier than in the other three versions.

- Stem and response:
  - Baseline version: Advertising on a notice board could turn out to be . . . (slow)
Task difficulty: ESL listening assessment

– Table format version: Empty cell in column ‘Advertise your-
self’ and row ‘Disadvantage’ (slow).

The item required candidates to provide a one-word response, a
common lexical item that is routinely taught at beginner level.
The word is in a stressed position in the text.

• Surrounding text:

– On the other hand, you simply have to wait until someone tele-
phones you, so this might be a rather slow means of getting a job.

The only cue for this item is ‘might be’ which is a paraphrase of
‘could turn out to be’ included in the stem. There was no lexical or
syntactic match of the stem with the text, a feature that has been
found to make items easier (Freedle and Kostin, 1996; Jensen et al.,
1997). If the listeners were looking for a lexical match to prepare
them for the necessary information – as they had in the previous
items in this test – they would have had to retain the cue from the
previous section of text where the lexical match occurred: ‘advertise
your services on the notice board yourself’. If they were relying on
this match, they would have been required to retain the information
over 54 words, thus increasing the cognitive load for that item. (In
this context, Jensen et al. [1997: 161] found that longer time between
tested information made items more difficult.) Other features of the
stem in this item were the complex grammar used (gerund and modal)
and the lack of recycling of the information. It could be hypothesized
that this combination of characteristics (complex grammar, lack of
lexical overlap in the immediate surrounding text, difficult paraphrase
of text to match the stem and the cognitive load of retaining infor-
mation from the initial cue over an extended text to match with the
stem) contributed to the difficulty of the item.

In contrast, the same item in the ‘table’ format version was
explicitly cued through the rhetorical structure of the text. The column
heading (‘advertise yourself’) matched the text ‘advertise your ser-
vice on the notice board yourself’ and the row heading
(‘disadvantage’) was clearly signalled in the text by ‘on the other
hand’ (see Figure 3). Listeners were not looking for a lexical match
in this version, but were relying on the regular rhetorical structure of
the text which consistently signalled the advantages first, followed by
the disadvantages. Candidates completing the table knew that the next
piece of information in this case would be a disadvantage. We may
speculate that this explicit signalling provided a clear macro-structure
for the listener, thus facilitating comprehension (Chaudron and
Richards, 1986).

The effects of some of the features described above can also be
noted in Item 8 of the ‘jobs’ task, ‘Jobs advertised on notice boards
‘Table’ item format

In the jobs task, a speaker describes the advantages and disadvantages of a number of different methods of looking for work.

Certificate in Spoken and Written English / Certificate III Assessment
Competency 3: Can demonstrate an understanding of an oral presentation

You will hear a talk about different ways of finding a job. Look at the questions and as you listen complete the table with the information you hear on the tape. Use only a few words. You will only hear the recording once.

<table>
<thead>
<tr>
<th>Advantage</th>
<th>Disadvantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ways of finding a job</td>
<td>Community group</td>
</tr>
<tr>
<td>Hear about jobs</td>
<td>1. Good for</td>
</tr>
<tr>
<td>Choices limited</td>
<td>2.</td>
</tr>
<tr>
<td></td>
<td>3.</td>
</tr>
<tr>
<td></td>
<td>4.</td>
</tr>
<tr>
<td></td>
<td>5.</td>
</tr>
<tr>
<td></td>
<td>6.</td>
</tr>
<tr>
<td></td>
<td>7.</td>
</tr>
<tr>
<td></td>
<td>8.</td>
</tr>
<tr>
<td></td>
<td>9.</td>
</tr>
<tr>
<td></td>
<td>10.</td>
</tr>
</tbody>
</table>

Figure 3 Response sheet for ‘table’ version of ‘jobs’ task

are usually . . . ’ (answer: badly paid/not permanent) (see Figure 1). This item was quite difficult in versions 2A, 2C and 2D (all using the baseline text version and item format, but with manipulated task conditions), despite the two-word verbatim response and simple vocabulary of the stem. It was most probably cued through a lexical match with ‘notice boards’, which occurs in the preceding part of the text and serves as a cue for both Item 7 and Item 8. To answer the test question correctly, listeners would have had to retain this cue over a 26-word segment of text before encountering the NI. As in the case of Item 10, the distance between the cue and the NI may have contributed to the difficulty of this item.

On the other hand, in the ‘table’ format version, the rhetorical structure of the text is somewhat more transparent and there is less time between the tested information and the response. The repeated cue ‘good for casual jobs/work’ provides candidates with the NI for Item 7 and alerts them that they are about to hear a disadvantage of notice boards, which they are required to write down in Item 8. This is clearly signalled by the macro-marker ‘but . . . ’ in the text, followed
Table 4  Speech rates for ‘dogs’ task

<table>
<thead>
<tr>
<th></th>
<th>Task 3A and 3B</th>
<th>Task 3C</th>
<th>Task 3D</th>
</tr>
</thead>
<tbody>
<tr>
<td>First minute</td>
<td>172 wpm</td>
<td>190 wpm</td>
<td>205 wpm</td>
</tr>
<tr>
<td>Second minute</td>
<td>157 wpm</td>
<td>186 wpm</td>
<td>200 wpm</td>
</tr>
<tr>
<td>Third minute</td>
<td>177 wpm</td>
<td>167 wpm</td>
<td>n/a</td>
</tr>
<tr>
<td>Fourth minute</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>Total number of minutes</td>
<td>3.30 mins</td>
<td>4.03 mins</td>
<td>2.53 mins</td>
</tr>
<tr>
<td>Total number of words</td>
<td>607 words</td>
<td>764 words</td>
<td>607 words</td>
</tr>
</tbody>
</table>

Notes: Average speech rate for Baseline version and item format version: 3A and 3B = 173 wpm (words per minute); Average speech rate for Dialogue version task 3C = 188 wpm; Average speech rate for Fast version task 3D = 210 wpm.

by the NI for Item 8. The response to this item involves simply a verbatim repeat of the words in the text. The presence of these factors may help to explain why Item 8 in this version was the easiest item in the ‘jobs’ task.

c Effects of speech rate: The items in the ‘faster’ version were also examined in order to investigate why Items 6 and 4 in the ‘dogs’ task should remain easier than the other items, despite the faster rate of delivery. To determine relative speech rates across the three versions, word per minute (wpm) counts were carried out on the ‘faster’ version, the baseline version, and the ‘dialogue’ version. Since these were quite long texts, wpm counts for each minute were calculated. Speech rates are set out in Table 4.

These counts suggest that there was quite a lot of variation in the speech rates throughout the texts. They are, nevertheless, still rough measures: a word per second count would give a truer picture of local fluctuations in speech rate and their impact on the difficulty of the corresponding item. The average speech rate for Task 3A and 3B is slower than Griffiths’ (1992) ‘average’ (188 wpm), whereas the faster speech rate is above the 200 wpm threshold, beyond which listening comprehension has been found to decline (Griffiths, 1990). However, despite the lower than average rates in Tasks 3A and 3B, there were still items in those versions (Items 1–5 and 8–10) that were more difficult than items in the faster version (Items 6 and 7). To explore why this might be the case, we conducted an analysis of the speech rates of the two participants in the ‘dialogue’ version, which revealed that they were using different rates. Moreover, the information containing the responses to the test questions was given by the faster of the two speakers. This may explain the fact that the ‘dialogue’ version was only marginally easier than the baseline version, despite the presence of other features (such as recycling and repetition of...
V Discussion

1 Interaction of task components

In this investigation we started out to look at the effect of particular task characteristics and conditions on the difficulty of listening texts, in the hope that we could identify a small number of ‘levelling variables’ (Dunkel et al., 1993) that could be used to regulate the difficulty of listening assessment tasks. However, it seems clear that there is a complex interaction between the different components of the task. In some cases, a particular item may remain ‘easy’, even in a ‘difficult’ version of a task as reflected in mean item difficulty estimates. This effect was observed in both the ‘faster’ version and the ‘table’ format version. Item 6, the easiest item, was still easier in the ‘faster’ version at 200 wpm.

Our analysis suggests that particular combinations of item characteristics appear either to accentuate or attenuate the effect on difficulty. In some cases, ‘easy’ or ‘difficult’ features may well cancel each other out; for example, the easy, high frequency, one word response in Item 10 in the ‘jobs’ task may have been made more difficult by the complex syntax of the stem, the weak match with the cue and a long retention time. Any conclusions regarding the effect of any single task or item characteristic on difficulty would therefore need to be carefully qualified in the light of what is known about its interaction with other variables. In this case, it could be said that one-word responses seem to be easier if there is a clear lexical match with the text. However, in addition, it would appear that the stem needs to be strongly cued in the text and the total retention time must be short. Similar interactions between item characteristics were identified by Jensen et al. (1997) who used hierarchical linear modelling to investigate the effects of text and item characteristics on item difficulty in an English for Academic Purposes listening test. They found, inter alia, a significant interaction between type of response and length of text: as the length of the text decreased, items requiring a verbatim response – as opposed to a nonverbatim response – became easier.

2 Methodological issues

In interpreting the results of this investigation, a number of methodological issues arose which affected the quality and quantity of the data.
that we obtained and, thus, the conclusions that we were able to draw. One of the main problems we encountered was in operationalizing the speech rate variable. In this context, research has identified the difficulties involved in defining speech rate and especially in deciding what constitutes a ‘fast’ as opposed to a ‘slow’ rate (Tauroza and Allison, 1990; Rubin, 1994). These difficulties are further compounded by the fact that speech rates may vary throughout a text. An average wpm count will not reflect these differences and therefore cannot provide any information that will help the researcher to understand the effect of speech rate at the level of the individual item. Since, as we have seen, speech rate may interact with other item-level variables, it could be argued that varying the task conditions by simply making a task ‘faster’ in terms of average wpm would not automatically result in an increase in difficulty unless these other factors are also taken into account.

Problems of operationalization also arose when we tried to examine the effects of the text type variable. We hypothesized that the ‘dialogue’ version of the ‘dogs’ task would be easier than the monologue version since it included more features of spoken language, recycling of information and greater redundancy. However, what we did not control here (and should have) was the speech rate variable. Since one of the participants in the dialogue was using a fast speech rate, and the majority of responses were spoken by this person, it is possible that the faster rate used in the dialogue version could have nullified any potentially ‘easifying’ effects of the other factors. Consequently, we were unable to draw any meaningful conclusions concerning the effect of text type. This highlights the need to try to specify and, where necessary, to quantify all variables as precisely as possible in studies of this kind. However, it also alerts us once again to the complexities of the interactions between the variables under investigation.

3 Implications for assessment task design

a Validity issues: This study has highlighted a number of problems involved in developing valid and reliable assessments of listening competencies in outcomes-based systems of assessment and reporting such as the CSWE. Foremost among these is the question of construct validity. Since the elements and performance criteria in the competency statement must identify specific language behaviour, the test-developer is placed in the position of having to describe the listening skills that contribute to test performance and then design a task that elicits these. However, a number of research studies have cast doubt on the extent to which separate skills can either be identified or
matched to particular listening items, since the interactive processing involved in listening may involve the simultaneous exercise of a range of skills (Buck, 1990; 1991b; 1994; Brindley, 1997). The descriptions of listening ability that traditionally figure in proficiency rating scales and competency statements must therefore be open to question, since there is no guarantee that a candidate who answers a test question correctly has used the designated skill(s) to arrive at the right answer (for further discussion see Alderson, 1991; Buck, 1991b; 1994; Brindley, 1997; 1998c). These findings weaken any validity claims that can be made concerning learners’ achievements of the specific elements of the competency in question.

b Reliability issues: These problems of construct validity are exacerbated by the fact that under normal classroom conditions it is unlikely that teachers would have time to develop a test of sufficient length to meet minimum standards of reliability, even for low stakes assessment (the sample assessment task for Competency 3 included in the CSWE documentation, for example, contains only three items, one of which requires extended writing; no marking key is supplied). These problems, combined with the variability across task characteristics and task conditions that we have found, suggest that reliable assessment would be very difficult to achieve under current conditions.

c Practical issues: defining competencies and standardizing tasks: The outcomes of this study point to the need for a more detailed and tightly defined specification of the listening competency investigated here and, by extension, of other competencies. If, for example, differences in features such as speed of delivery and item format can impact on item and task difficulty and thus lead to potentially unfair decisions regarding competency attainment, these features need to be spelled out more explicitly in future versions of the CSWE. Item characteristic variables that appear to have more predictable effects (such as amount of lexical overlap, verbatim vs nonverbatim responses and repetitions, and the speech rate of the surrounding text) could also be built into task design as a way of controlling item difficulty (Jensen et al., 1997: 163).

Including task or item characteristic variables in the competency specifications, however, creates a dilemma in outcomes-based systems where teachers are also the task developers. For example, it is highly unlikely that a teacher in the real world would be in a position to construct a listening assessment task that incorporated speech at a given rate. Given the difficulty of controlling this and the many other factors involved in listening comprehension, there would appear to be a case for creating a set of standard tasks, rather than leaving task
design up to individual teachers. Item specifications, however, would need to include guidance to writers on the key variables that were most likely to influence difficulty. The tasks that were developed could then be piloted, calibrated for difficulty and used as ‘anchors’ against which new tasks could be compared as they were progressively created (Brindley, 2001). At the same time, minimum standards for the achievement of each listening task would also need to be set, based on examination of item response patterns for the task in question.

VI Conclusions

In this study we set out with the practical goal of determining whether changes in task characteristics and task conditions in competency-based listening tasks would result in differences in test performance. Knowing which variables were likely to affect test scores would, we hoped, provide a basis for controlling task difficulty and thus for making tasks more comparable in the interests of fairness. However, the complexities of the interactions between task characteristics, item characteristics and candidate responses that we have encountered in this study suggest that simply adjusting one task-level variable will not automatically make the task easier or more difficult.

Given these complexities, we must conclude that the rather simplistic notion of ‘difficulty’ as reflected in item difficulty statistics is of limited usefulness in understanding what happens when an individual candidate interacts with an individual item. As a number of participants in a recent discussion on the International Language Testing Association discussion list, LTEST-L, noted, the question of ‘what makes an item difficult?’ may be the wrong question to be asking. Lyle Bachman (posting to LTEST-L, 19 February 2000) sums up the problem succinctly:

As soon as one considers what makes items difficult, one immediately realizes that difficulty isn’t a reasonable question at all. A given task or item is differentially difficult for different test takers and a given test taker will find different tasks differentially difficult. Ergo, difficulty is not a separate quality at all, but rather a function of the interaction between task characteristics and test taker characteristics. When we design a test, we can specify the task characteristics, and describe the characteristics of the test takers, but getting at the interaction is the rub.

Seen from this perspective, the design of listening tasks that are comparable across different administrations becomes a formidable challenge. If, as Buck (1994: 164) suggests, ‘performance on each test item by each test-taker is a unique cognitive event’, then task design will require not only a much more detailed specification of task
characteristics and conditions (Bachman and Palmer, 1996), but also it will need to be based on a much better understanding of the interactions between text, task and learner variables. To this end, a good deal of further work will need to go into building models of listening test performance that incorporate a wide range of overlapping difficulty components and exploring their effects on test scores (Dunkel et al., 1993; Jensen et al., 1997; Buck and Tatsuoka, 1998; Norris et al., 1998). At the same time, continued research using introspective methods will be required to uncover the complex cognitive skills that underlie item responses and to determine how these are mobilized in particular listening tasks (Buck, 1990; 1991; 1994; Buck and Tatsuoka, 1998; Wu, 1998). In the meantime, the goal of designing valid and reliable tasks for assessing listening competencies will remain elusive.

Acknowledgements

The research reported in this article was carried out under the Special Projects Research Program funded by the Commonwealth Department of Immigration and Multicultural and Indigenous Affairs through the AMEP Research Centre, a consortium of the National Centre for English Language Teaching and Research, Macquarie University and the Institute for Education, La Trobe University. We would like to express our gratitude to Steven Ross for his advice on the study design, to Helen Price for her assistance with data collection and analysis and to the AMEP teachers who gave up their time to participate in the development, trialling and administration of the assessment tasks.

VII References


Linguistics and the American Association of Teachers of Spanish and Portuguese, Cancun, Mexico, August 9–13, 1992. ED 358737.


**Christie, J. and Delaruelle, S.** 1997: *Assessment and moderation*. Sydney: National Centre for English Language Teaching and Research, Macquarie University.

**Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. and McNamara,**


