ResearchNotes

Editorial Notes

Welcome to issue 13 of *Research Notes*, our quarterly publication reporting on matters relating to research, test development and validation within Cambridge ESOL.

The focus of this issue is on speaking, a component of all of Cambridge ESOL's language testing products. In this issue we approach the concept of speaking in various ways: from a consideration of our construct of speaking; through developing tests and ways of assessing those tests, through to the analysis of how candidates and examiners perform in speaking tests.

In the opening article Lynda Taylor describes Cambridge ESOL's approach to the assessment of speaking, including a brief history of the Cambridge speaking tests and our view of the construct of speaking. Lynda also summarises the features of test format and task design, test conduct and assessment that make Cambridge ESOL's speaking tests unique. Alongside considerations of what form our speaking tests take, Lynda describes how Cambridge ESOL is investigating the issues surrounding speaking tests that include analysis of test-taker and examiner talk and analyses of criteria and rating scales.

Following on from this, Michael McCarthy and Ron Carter (University of Nottingham) investigate what vocabulary is used most frequently in day-to-day spoken interaction. This consideration of one aspect of speaking describes the range of vocabulary produced by native speakers and is relevant to our own testing of speaking. Cambridge ESOL is currently developing its own collection of speaking tests and is collaborating with McCarthy and colleagues on the transcription and analysis of some of our learner English data.

Angela ffrench describes in detail how a new set of assessment criteria were developed for the Certificate of Proficiency in English and reflects on the impact of these revised criteria for other Main Suite examinations. In the following article Stuart Shaw reports on a recent administration of the CELS speaking test in terms of how the candidates and examiners behaved in this session. His article on on-line examiner reliability will appear in a future issue. David Booth evaluates the success of the revised Business English Certificate (BEC) speaking tests, describing how the delivery of BEC speaking tests are monitored, covering the scoring of the test, examiner and candidate choices.

Cambridge ESOL staff have recently attended a range of conferences, some of which are reported on in the Conference Reports section. Paul Seddon and Trish Burrow report on a pre-conference event on computer based tests and a presentation on determining suitable tasks for young learners, topics covered in *Research Notes* 7 (young learners) and 12 (technology). Lynda Taylor reports on the AAAL and TESOL conferences which were held in March this year.

Offprints of eighty *Research Notes* articles are now available to view or download from the Cambridge ESOL website. Further details can be found at the end of this issue.

We can now announce that Dr Michael Milanovic has been appointed to the role of Chief Executive of Cambridge ESOL with immediate effect. Mike, whose academic background is in Applied Linguistics and Language Testing joined UCLES in 1989 following a career in language teaching and testing, first in France then Hong Kong where he worked mainly with the British Council, the Hong Kong Examinations Authority and the City Polytechnic. After starting at UCLES as Head of the EFL Evaluation Unit, Mike in 1998 assumed overall responsibility for EFL Operations and Assessment following the introduction of business streams.

Contents

The URL for reading/downloading issues of *Research Notes* is:
http://www.CambridgeESOL.org/researchnotes

The URL for subscribing to *Research Notes* is:
http://www.CambridgeESOL.org/researchnotes/inform.cfm

Published August 2003

The Cambridge approach to speaking assessment

LYNDA TAYLOR, RESEARCH AND VALIDATION GROUP

Introduction

Direct[1] tests of speaking (and writing) have always been standard practice in Cambridge examinations for assessing both first (L1) and second (L2) language proficiency and Cambridge ESOL has a long experience of 'direct' speaking assessment. Although some writers have claimed that performance tests date back only to the middle of the 20th century (McNamara 1996, Lowe 1988), it is worth noting that the very first UCLES English language proficiency examination – the *Certificate of Proficiency in English (CPE),* introduced in 1913 – included a compulsory oral component: candidates faced half an hour of reading aloud and conversation with the oral examiner plus half an hour of dictation; the result was a speaking test which lasted a whole hour. Almost a century ago the direct assessment of spoken language proficiency was considered by Cambridge examinations to be very important; and since 1913, each new Cambridge ESOL examination has included a face-to-face speaking test as an integral part of the overall proficiency assessment. Today this feature is generally recognised as one of the strengths of our approach to testing learners' English.

The construct of L2 spoken language proficiency

It is probably true to say that Cambridge ESOL speaking tests have always reflected a view of speaking ability which involves multiple competences (e.g. lexico-grammatical knowledge, phonological control, pragmatic awareness); all these factors were represented either explicitly or implicitly in the test format and the assessment criteria over many decades. Little has changed in this respect, although today the underlying construct of spoken language proficiency operationalised in our ESOL speaking tests takes account of more modern, cognitive descriptions of the speech production process (e.g. Levelt 1989; Garman 1990). Such views hold that the proficient L2 speaker possesses the following competences:

(a) a wide repertoire of lexis and grammar to enable flexible, appropriate, precise construction of utterances in 'real time' (the *knowledge* factor);

(b) a set of established procedures for pronunciation and lexico-grammar, and a set of established 'chunks' of language, all of which will enable fluent performance with 'on-line' planning reduced to acceptable amounts and timing (the *processing* factor).

In addition, spoken language production tends to be based in social interaction, to be purposeful and goal-oriented within a specific context; and, while it is capable of being routine and predictable, it also has the capacity for relative creativity and unpredictability. Research in recent years has highlighted various features that are characteristic of more or less proficient oral performances (see Tonkyn and Wilson 2003 for a list of useful studies which can help oral test designers identify theoretically relevant and helpfully discriminating features of performance).

Features of test format and task design

Our current understanding of the nature of L2 spoken language proficiency directly informs features of test format and task design in the Cambridge ESOL speaking tests. Concern for authenticity of test content and tasks and the relationship between the "input" and the expected response or "output" is an important feature of content validation; the authenticity of the tasks and materials in the Cambridge tests is often referred to as a major strength of the approach. Test content must be designed to provide sufficient evidence of the underlying abilities (i.e. construct) through the way the test taker responds to this input. The authenticity of test content and the authenticity of the candidate's interaction with that content are important considerations in achieving high validity; some of the more familiar features of Cambridge speaking tests directly reflect this concern for authenticity:

- the pairing of candidates, where possible (to allow for a more varied sample of interaction, i.e. candidate-candidate as well as candidate-examiner);
- the multi-part test format (to allow for different patterns of spoken interaction, i.e. question and answer, uninterrupted long turn, discussion);
- the use of analytical and global criteria (to allow for a focus on overall discourse performance as well as on specific features such as lexical range, grammatical accuracy and phonological control).

Features of test conduct and assessment

As well as informing speaking test format and task design, the underlying construct of spoken language ability also shapes the choice and definition of assessment criteria; and it is principles of good measurement which determine other key features such as:

- the pairing of examiners (with one acting as participant-interlocutor and one as observer-assessor – but both providing an assessment of performance, i.e. multiple observations);

[1] As opposed to 'indirect' methods of testing speaking in which proficiency is assessed using a pre-recorded cassette or even a written test.

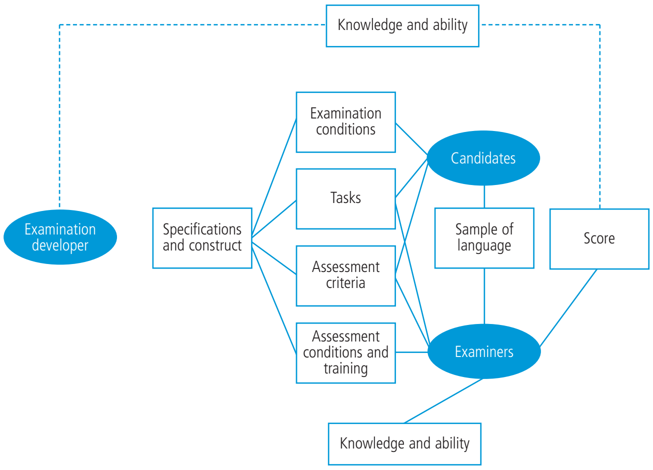
- the use of an interlocutor frame (to guide the management of the test and to ensure that all candidates receive similar, standardised input in terms of test format and timing);
- the implementation of a comprehensive oral examiner training/standardisation programme (to increase the reliability of subjectively judged ratings and provide a common standard and meaning for such judgements – see Alderson 1991).

Previous issues of *Research Notes* have regularly discussed some of the key features noted above in relation to specific Cambridge ESOL speaking tests (see various articles by Taylor).

Investigating the issues surrounding speaking tests

Direct testing of spoken language proficiency is a complex endeavour due to the many different variables or 'facets' which interact in performance assessment. Milanovic and Saville (1996) provide a useful overview of these variables and suggest a conceptual framework for setting out different avenues of research (see Figure 1).

Figure 1: Milanovic and Saville (1996)



This framework was influential in projects to develop/revise the speaking components of the Cambridge ESOL examinations during the 1990s – including the development of KET and CAE, as well as revisions to PET, FCE (see Saville and Hargreaves 1999) and, more recently CPE (Weir and Milanovic 2003). The framework was first presented at LTRC 1993 and is one of the earliest and most comprehensive of these models (see also Kenyon 1995 and McNamara 1996). It is valuable in that it highlights the many factors (or facets) which must be considered when designing a speaking test from which particular inferences are to be drawn; all of the factors represented in the model pose potential threats to the reliability and validity of these inferences and they need to be investigated as part of an ongoing research and validation programme for speaking tests.

Interestingly, Cambridge ESOL's experience of researching the complex issues which surround face-to-face speaking assessment dates back much further than just the past decade. As early as 1945, John Roach (then Assistant Secretary at UCLES) produced a report entitled *Some Problems of Oral Examinations in Modern Languages: An Experimental Approach Based on the Cambridge Examinations in English for Foreign Students*. Roach was particularly interested in how to describe levels of L2 speaking performance, and how to standardise oral examiners so that they rate candidates in a fair and consistent manner. Nowadays much of our research effort at Cambridge still goes into analysing the talk produced by candidates in our speaking tests in order to describe as usefully as possible for test users (i.e. candidates, teachers, employers) what it means to have a certain level of spoken language ability in English.

Investigating test-taker talk

Various approaches are used to investigate test-taker talk: for example, detailed transcription analysis allows us to investigate aspects of grammatical, lexical and discourse control (see examples of this type of research in Lazaraton 2002); the use of observational checklists enables us to study the range and frequency of spoken language functions which can be elicited by different task types. Both types of analysis help us to confirm the key components of L2 spoken language performance and provide the basis on which to create valid assessment criteria for making judgements about the quality of a learner's English language proficiency. By studying speech samples at different proficiency levels we can build performance descriptors which are used by trained and standardised oral examiners to make those judgements and produce reliable speaking test scores for individual test-takers, as well as more user-oriented level descriptions such as those which make up the Cambridge Common Scale for Speaking (see Angela ffrench's article on page 8).

Transcription and observational analyses also help us to design effective tasks for use in our speaking tests, and to answer questions such as: will this task generate enough language for assessment purposes? will it produce the right sort of language for this level? In this way we can confirm that a new task type will be suitable for use or that a revised test format is indeed functioning as intended (see *Research Notes* 2, 3 and 11).

Investigating examiner talk

Test-takers are of course not the only people who produce talk during a speaking test. The oral examiner also produces spoken output and this too has been the subject of ongoing research investigation over many years. Various studies have highlighted the problems of variation in examiner talk across different test-takers and the extent to which this can affect the opportunity candidates are given to speak, the language they produce and the score they receive. The results of such studies have confirmed the value of using a standardised script or 'interlocutor frame' in our speaking tests.

Our research into oral examiner talk has also led to an approach for assessing oral examiner performance based upon an Oral Examiner Monitoring Checklist. Oral examiners are routinely monitored during live examining by their team leader – a senior and more experienced oral examiner; the checklist is used to monitor systematically the quality of an oral examiner's conduct and assessment of the speaking test event. Analysis of the data gathered is then used to provide feedback to examiners and trainers on how examining techniques might be improved and is an important measure of the extent to which our speaking tests are standardised worldwide.

Investigating criteria and rating scales

The development of speaking assessment criteria and rating scales (or band descriptors) is clearly an important focus of our research and Angela ffrench's article on page 8 provides a comprehensive description of work in this area.

Investigating test scores

By the end of a speaking test, each test-taker has received a set of marks which reflects the quality of their spoken language performance. Collection and analysis of marks means we can analyse the score data statistically and answer a number of interesting questions relating to the performance of not only the candidates but also the oral examiners, the test materials, and the assessment criteria and scales. Results from such analyses help to provide evidence in support of the assumptions which underpin a test's design: that examiners are using the measurement scales as intended and are behaving in a consistent manner; that the multiple versions of speaking test material needed for security reasons are comparable in difficulty; that the assessment criteria are a valid reflection of what constitutes spoken language proficiency. Analyses of this type are routinely carried out following live examination sessions so that the information can feed directly back into the ongoing test development, production and revision cycle for all our speaking tests (see articles by Stuart Shaw and David Booth on pages 16 and 19).

Conclusion

The last few years have brought applied linguists and language testers a much clearer understanding of the nature of L2 spoken language proficiency and of the many different research avenues which are open to us. New and sophisticated methodologies – both qualitative and quantitative – are now available to help us in the complex business of investigating speaking tests. Advances in the area of corpus linguistics are especially exciting and at Cambridge we have recently started to build a small corpus of spoken learner data using audio-recordings of our speaking tests; over time, analysis of this corpus should provide us with rich additional insights into the nature of spoken language proficiency across different levels (preliminary, intermediate, advanced) and across different linguistic domains (general, business, academic). The article by Michael McCarthy and Ron Carter on page 5 is a good indication of recent progress in this area and of future promise.

As this special issue demonstrates, almost a full century after introducing its first face-to-face speaking test, Cambridge ESOL remains at the cutting edge of the direct assessment of L2 spoken language proficiency.

References and Further Reading

Alderson, J C (1991): Bands and scores, in Alderson J C and North B (eds): *Language testing in the 1990s: the communicative legacy*, London: Macmillan.

Garman, M (1990): *Psycholinguistics*, Cambridge: Cambridge University Press.

Kenyon, D (1995): An investigation of the validity of task demands on performance-based tests of oral proficiency, in Kunnan, A J (ed): *Validation in language assessment: selected papers from the 17th Language Testing Research Colloquium*, Long Beach, Mahwah, NJ, Lawrence Erlbaum Associates Publishers, 19–40.

Lazaraton, A (2002): *A qualitative approach to the validation of oral language tests*, Studies in Language Testing 14, Cambridge: UCLES/Cambridge University Press.

Levelt, W (1989): *Speaking: from intention to articulation*, Cambridge, Mass: MIT Press.

Lowe, P (1988): The unassimilated history, in Lowe P and Stansfield C W (eds): *Second language proficiency assessment: current issues*, Englewood Cliffs, NJ: Prentice Hall Regents, 11–51.

McNamara, T (1996): *Measuring second language performance*, London: Longman.

Milanovic, M and Saville, N (1996): Introduction in *Performance testing, cognition and assessment: selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem*, Studies in Language Testing 3, Cambridge: UCLES/Cambridge University Press, 1–17.

Roach, J (1945): *Some Problems of Oral Examinations in Modern Languages: An Experimental Approach Based on the Cambridge Examinations in English for Foreign Students.*

Saville, N and Hargreaves, P (1999): Assessing speaking in the revised FCE, *English Language Teaching Journal*, 53/1, 42–51.

Tonkyn, A and Wilson, J (2003): *Revising the IELTS Speaking Test*, in the proceedings of the BALEAP 2001 Annual Conference, Strathclyde.

Weir, C and Milanovic, M (2003): *Continuity and innovation: revising the Cambridge Proficiency in English Examination 1913–2002*, Studies in Language Testing 15, Cambridge: UCLES/Cambridge University Press.

What constitutes a basic spoken vocabulary?

MICHAEL MCCARTHY AND RONALD CARTER, UNIVERSITY OF NOTTINGHAM

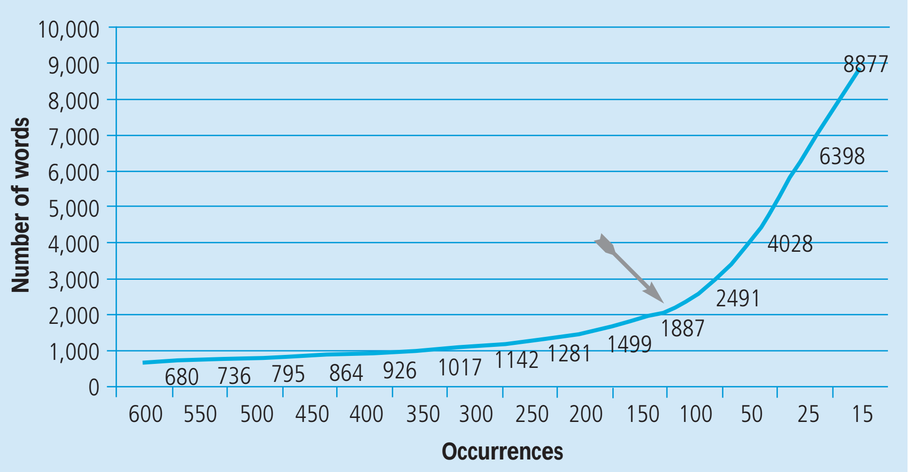
Introduction

In the last 20 years or so, corpus linguists have been able to offer computerised frequency counts based on written and, more recently, spoken corpora. In this article we look at frequency in the 5-million word CANCODE spoken corpus (see McCarthy 1998). CANCODE stands for Cambridge and Nottingham Corpus of Discourse in English. The corpus was established at the Department of English Studies, University of Nottingham, UK, and is funded by Cambridge University Press, with whom the sole copyright resides. We also look at the spoken element of the British National Corpus (BNC) (see Rundell 1995a and b; Leech et al 2001). The spoken BNC amounts to 10 million words, and the corpus is in the public domain. Frequency statistics from the BNC are available in Leech et al (2001). Using such resources it is possible to obtain at least some answers to the question: what vocabulary is used most frequently in day-to-day spoken interaction?

How big is a basic vocabulary?

There is no easy answer to this question, except to say that, in frequency counts, there is usually a point where frequency drops off rather sharply, from extremely high frequency, hard-working words to words that occur very infrequently. In other words, frequencies do not decline at a regular rate, but usually have a point where there is a sudden change to low frequency. This applies to both spoken and written corpora. The point where high frequency suddenly drops to low can be seen as a boundary between the core and the rest, though that point might be expected to vary a little from corpus to corpus. Figure 1 shows how frequency drops off in a 5-million word spoken sample of the BNC. The horizontal axis shows frequency bands (i.e. 15 indicates a band of words occurring 15 times in the corpus, 400 = a band of words occurring 400 times, etc.). The vertical axis shows how many words in the corpus actually occur at those bands (e.g. around 2500 words occur 100 times).

Figure 1: Frequency distribution: 5 million words BNC spoken



Round about 2000 words down in the frequency ratings (indicated by an arrow), the graph begins to rise very steeply, with a marked increase in the number of words that occur less than 100 times, such that almost 9000 words are occurring 15 times. Even at an occurrence level of 50, there are more than 4000 words. We can conclude that words occurring 100 times or more in the spoken corpus belong to some sort of heavy-duty core vocabulary, which amounts to about 2000 words. It is reasonable to suppose, therefore, that a round-figure pedagogical target of the first 2000 words will safely cover the everyday spoken core with some margin for error.

In the case of written data, the same phenomenon occurs (i.e. a similar shape of graph), but the number of words in the core is greater. We see a similar abrupt change from the core, high-frequency words to a huge number of low frequency items, but that change occurs at over 3000 words, not 2000. This is not surprising, since lexical density and variation is greater in written than in spoken texts.

Some observations on the spoken core

Table 1 lists the words that occur in excess of 1,000 times per million words in the BNC and in CANCODE, and thus perform heavy duty.

The BNC and CANCODE are remarkably consistent on the top 100 words, suggesting a good level of reliability for the figures. However, questions arise as to the place of many of these items in a ‘vocabulary’ list. The first 100 include articles, pronouns, auxiliary verbs, demonstratives, basic conjunctions, etc. The types of meaning they convey are traditionally considered to be grammatical rather than lexical. Another problem raised by the top 100 list is that of fixed phrases, or ‘chunks’ extending over more than one word. Word #31 (*know*) and word #78 (*mean*) are so frequent mainly because of their collocation with *you* and *I*, in the formulaic phrases *you know*, and *I mean*.

All in all, the top 100 BNC spoken list

shows that arriving at the basic vocabulary is not just a matter of instructing the computer to list the most frequent forms, and considerable analytical work is necessary to refine the raw data. The computer does not know what a vocabulary item is. Nonetheless, the top 2000 word list is an invaluable starting point, for a good many reasons, not least because clear basic meaning categories emerge from it. Those basic categories are what the rest of this article is about. If, on the basis of general professional consensus, we exclude as a category anything up to 200 grammar/functional word-forms, the remainder of the 2000 word list falls into roughly nine types of item. These are not presented in any prioritised order, and all may be considered equally important.

Modal items

Modal items carry meanings referring to degrees of certainty or necessity. The 2000 list includes the modal verbs (*can, could, will, should*, etc.), but the list also contains other high frequency items carrying related meanings. These include the verbs *look, seem* and *sound*, the adjectives *possible* and *certain* and the adverbs *maybe, definitely, probably* and *apparently*. The spoken list offers compelling evidence of the ubiquity of modal items in everyday communication, beyond the well-trodden core modal verbs.

Delexical verbs

This category embraces high-frequency verbs such as *do, make, take* and *get*. They are called delexical because of their low lexical content and the fact that their meanings are normally derived from the words they co-occur with (e.g. *make a mistake, make dinner*). However, those collocating words may often be of relatively low frequency (e.g. *get a degree, get involved, make an appointment*), or may be combinations with high-frequency particles generating semantically opaque phrasal verbs (e.g. *get round to doing something, take over from someone*).

Interactive markers

There are a number of items which represent speakers' attitudes and stance. These are central to communicative well-being and to maintaining social relations. They are not a luxury, and it is hard to conceive of anything but the most sterile survival-level communication occurring without them. The words include *just, whatever, thing(s), actually, basically, hopefully, really, pretty, quite, literally*. The interactive words may variously soften or make indirect potentially face-threatening utterances, purposely make things vague or fuzzy in the conversation, or intensify and emphasise one's stance.

Table 1: 100 most frequent items, total spoken segment (10 million words), BNC

	Word	Frequency per 1m words		Word	Frequency per 1m words		Word	Frequency per 1m words
1	the	39605	35	got	5025	68	two	2710
2	I	29448	36	‘ve	4735	69	said	2685
3	you	25957	37	not	4693	70	one	2532
4	and	25210	38	are	4663	71	m	2512
5	it	24508	39	if	4544	72	see	2507
6	a	18637	40	with	4446	73	me	2444
7	’s	17677	41	no	4388	74	very	2373
8	to	14912	42	‘re	4255	75	out	2316
9	of	14550	43	she	4136	76	my	2278
10	that	14252	44	at	4115	77	when	2255
11	-n’t	12212	45	there	4067	78	mean	2250
12	in	11609	46	think	3977	79	right	2209
13	we	10448	47	yes	3840	80	which	2208
14	is	10164	48	just	3820	81	from	2178
15	do	9594	49	all	3644	82	going	2174
16	they	9333	50	can	3588	83	say	2116
17	er	8542	51	then	3474	84	been	2082
18	was	8097	52	get	3464	85	people	2063
19	yeah	7890	53	did	3368	86	because	2039
20	have	7488	54	or	3357	87	some	1986
21	what	7313	55	would	3278	88	could	1949
22	he	7277	56	mm	3163	89	will	1890
23	that	7246	57	them	3126	90	how	1888
24	to	6950	58	'll	3066	91	on	1849
25	but	6366	59	one	3034	92	an	1846
26	for	6239	60	there	2894	93	time	1819
27	erm	6029	61	up	2891	94	who	1780
28	be	5790	62	go	2885	95	want	1776
29	on	5659	63	now	2864	96	like	1762
30	this	5627	64	your	2859	97	come	1737
31	know	5550	65	had	2835	98	really	1727
32	well	5310	66	were	2749	99	three	1721
33	so	5067	67	about	2730	100	by	1663
34	oh	5052						

Discourse markers

Discourse markers organise and monitor the talk. A range of such items occur in the top 2000 most frequent forms and combinations, including *I mean, right, well, so, good, you know, anyway*. Their functions include marking openings and closings, returns to diverted or interrupted talk, signalling topic boundaries and so on. They are, like the interactive words, an important feature of the interpersonal stratum of discourse. The absence of discourse markers in the talk of an individual leaves him/her potentially disempowered and at risk of becoming a second-class participant in the conversation.

Deictic words

Deictic words relate the speaker to the world in relative terms of time and space. The most obvious examples are words such as *this* and *that*, where ‘this box’ for the speaker may be ‘that box’ for a remotely placed listener, or the speaker's *here* might be *here* or *there* for the listener, depending on where each person is relative to each other. The 2000 list contains words with deictic meanings such as *now, then, ago, away, front, side* and the extremely frequent *back* (as the opposite of *front*, but mostly meaning 'returned from another place').

Basic nouns

In the 2000 list we find a wide range of nouns of very general, non-concrete and concrete meanings, such as *person, problem, life, noise, situation, sort, trouble, family, kids, room, car, school, door, water, house, TV, ticket*, along with the names of days, months, colours, body-parts, kinship terms, other general time and place nouns such as the names of the four seasons, the points of the compass, and nouns denoting basic activities and events such as *trip* and *breakfast*. These nouns, because of their general meanings, have wide communicative coverage. *Trip*, for example, can substitute for lower frequency items such as *voyage, flight, drive*, etc. In terms of everyday categories, there is a degree of unevenness. In the names of the four seasons in CANCODE, *summer* is three times more frequent than *winter*, and four times more frequent than *spring*, with *autumn* trailing behind at ten times less frequent than *summer* and outside of the top 2000 list. Pedagogical decisions may override such awkward but fascinating statistics. However, some closed sets are large (e.g. all the possible body parts, or the names of all countries in the world), and in such cases frequency lists are helpful for establishing priorities.

Basic adjectives

In this class there appear a number of adjectives for everyday positive and negative evaluations. These include *lovely, nice, different, good, bad, horrible, terrible*. Basic adjectives (and basic adverbs) often occur as response tokens (speaker A says 'See you at five', speaker B says 'fine/great/good/lovely'). *Great, good, fine, wonderful, excellent, lovely*, etc. occur very frequently in this function. These items make the difference between a respondent who repeatedly responds with an impoverished range of vocalisations or the constant use of *yes* and/or *no* and one who sounds engaged, interested and interesting.

Basic adverbs

Many time adverbs are of extremely high frequency, such as *today, yesterday, tomorrow, eventually, finally*, as are adverbs of frequency and habituality, such as *usually, normally, generally*, and of manner and degree, such as *quickly* (but not *slowly*, which comes in at word #2685), *suddenly, fast, totally, especially*. This class of word is fairly straightforward, but some prepositional phrase adverbials are also extremely frequent, such as *in the end*, and *at the moment*, which occur 205 and 626 times, respectively, in CANCODE. Once again, the single word-form list often hides the frequency of phrasal combinations (see McCarthy and Carter, in press).

Basic verbs

Beyond the delexical verbs, there are verbs denoting everyday activity, such as *sit, give, say, leave, stop, help, feel, put, listen, explain, love, eat*. It is worth noting the distribution of particular tense/aspect forms. Of the 14,682 occurrences of the forms of SAY (i.e. *say, says, saying, said*) in CANCODE, 5416 of these (36.8%) are the past tense *said*, owing to the high frequency of speech reports. Such differences may be important in elementary level pedagogy, where vocabulary growth often outstrips grammatical knowledge, and a past form might need to be introduced even though familiarity with the past tense in general may be low.

Conclusion

With spoken data, there is a core vocabulary based around the 1500–2000 most frequent words, a vocabulary that does very hard work in day-to-day communication. Written data has a larger core. However, raw lists of items need careful evaluation and further observations of the corpus itself before an elementary-level vocabulary syllabus can be established. Not least of the problems is that of widely differing frequencies within sets of items that seem, intuitively, to form useful families for language learning and testing purposes. Equally, the list needs to take account of collocations and phrasal items, as in the case of delexical verbs, discourse markers and basic adverbs. But the list can also be very useful in suggesting priorities for the grading of closed sets consisting of large numbers of items (e.g. the human body parts). Corpus statistics take us a considerable way from what intuition and conventional practice alone can provide, but the one should not exist without the other.

References and Further Reading

Leech, G, Rayson, P and Wilson, A (2001): *Word Frequencies in Written and Spoken English*, London: Longman.

McCarthy, M (1998): *Spoken Language and Applied Linguistics*, Cambridge: Cambridge University Press.

McCarthy, M and Carter, R (in press): This that and the other: Multi-word clusters in spoken English as visible patterns of interaction, *Teanga*. Special issue on corpus and language variation.

Rundell, M (1995a): The BNC: A spoken corpus, *Modern English Teacher*, 4/2, 13–15.

– (1995b): The word on the street, *English Today*, 11/3, 29–35.

The development of a set of assessment criteria for Speaking Tests

ANGELA FFRENCH, MAIN SUITE GROUP

This article charts the development of a set of assessment criteria for the Cambridge ESOL Certificate of Proficiency and its impact on the assessment criteria used for other examinations in the Cambridge ESOL Main Suite.

Introduction

In recent years, Cambridge ESOL has been developing a Common Scale of Assessment for Speaking for its Main Suite examinations, replacing assessment criteria which had been previously used. The Common Scale for Speaking project was developed in two stages. Stage One focussed on four of the examinations in the Cambridge ESOL Main Suite: Key English Test (KET), Preliminary English Test (PET), First Certificate in English (FCE) and Certificate in Advanced English (CAE), which are placed at Levels 1–4 on the Association of Language Testers in Europe (ALTE) Framework. Stage Two focussed on the revised Certificate of Proficiency in English (CPE) at ALTE level 5.

Stage One

Since the introduction of the Common Scale for Speaking in 1996, observations on the operational use of the criteria have been fed back via the Team Leader (TL) System (Taylor 2000), and regular analysis of marks awarded by Oral Examiners (OEs) during live examining sessions has been carried out. This information has helped to identify areas where OEs might be experiencing difficulty in separating out the different scales and/or being able to select an appropriate mark for a particular scale.

KET OEs apply a Global Achievement Scale to each of the two parts of the KET Speaking Test. This scale contains all the analytical elements of the criteria used for the Main Suite speaking tests at levels 2–4.

PET, FCE and CAE assessment criteria comprise four analytical scales (Grammar and Vocabulary, Discourse Management, Pronunciation, Interactive Communication), which are applied to the candidate's performance across the whole test by an OE who takes the role of assessor, and one global scale (Global Achievement), which takes an holistic view of the candidate's performance and is applied by an OE who takes the role of interlocutor. There are nine mark bands labelled 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0 and descriptors are attached to the bands of 1.0, 3.0 and 5.0. The band of 3.0 is seen to represent adequacy at a given level; 5.0 represents top of the range; 1.0 is seen to be an inadequate performance. A mark of 0 could also be awarded if a candidate failed to provide sufficient language for assessment.

The previous CPE assessment criteria comprised six analytical scales: Fluency, Grammatical Accuracy, Pronunciation (Prosodic Features), Pronunciation (Individual Sounds), Interactive Communication, Vocabulary Resource. Descriptors were attached to each of the marks for each of the scales and marks were awarded by one OE.

Through trialling and subsequent application of the Common Scale model, it was found that:

- descriptors for three of the nine bands, across the four analytical scales were sufficient to provide OEs with the information required to make their judgements. Candidates' performances may fit the exact wording of a descriptor, but there are many instances when the performance has elements of the description attached, say, to the 5.0 band and elements which are reflected in the wording of the 3.0 band. The OEs judgement is based on the degree to which the performance fits the descriptors;
- capturing independent assessments from two assessors was seen, by both examiners and candidates, to be fairer than assessment by a single examiner.

Stage Two

It was decided that the project to revise the assessment criteria for the revised CPE Speaking Test should take the Common Scale model as its starting point and draw on both operational observations and statistical data through trialling.

In order to arrive at a final operational set of assessment scales for the Revised CPE Speaking Test, a number of developmental stages were envisaged. The first stage (Phase 1) involved producing draft assessment scales and modifying these until it was felt they were ready to be trialled with UK TLs (Phase 2). As a result of this trial, the scales were modified (Phase 3) and this revised version was trialled with Senior Team Leaders (STLs) in Phase 4. Further modifications were made to the assessment scales (Phase 5) and this was followed by an exercise which compared the revised scales with the previous CPE assessment scales (Phase 6). Phase 7 looked at the impact of the revised CPE assessment scales on the PET, FCE and CAE assessment scales. By Phase 8, the assessment scales were ready to be trialled, with 11 STLs from around the world assessing the performances of 24 CPE level students. Finally, in Phase 9 the Global Achievement Scale was picked up and developed in line with Global Achievement scales for PET, FCE and CAE.

PHASE 1: INITIAL DRAFT (APRIL 2000)

At the start of the project, a specialist in the field of pronunciation, language testing and development of assessment criteria, was invited to join the Cambridge ESOL team responsible for the development of the assessment criteria. The team felt it was important to consider CPE in relation to the other Main Suite examinations and the Common Scale, but also to acknowledge differences where appropriate. Initial discussions concluded that it was feasible to separate vocabulary from grammar at CPE level and that (unlike the previous criteria) there should be only one scale for pronunciation.

When the Common Scale of Assessment was first developed, it was felt appropriate for the 'range' of grammar to be linked to the extent and coherence of a candidate's contribution. However, OEs reported difficulty in separating the appropriate use of a range of structures from their accurate application. Also, it seemed appropriate to include in Discourse Management the notion of relevance of the contribution, which had not previously been focussed upon. Therefore, it was agreed that 'range of grammar' should be removed from the scale of Discourse Management and repositioned with grammatical accuracy under the heading of Grammatical Resource.

The initial working document proposed a set of six scales, five analytical and one global: Grammatical Resource, Lexical Resource, Discourse Management, Pronunciation, Interactive Communication, and Global Achievement.

This preliminary work focussed on three areas:

1. Clarification of the analytical assessment scales – Explanations of Criteria. These identified the different foci of each scale, e.g. Grammatical Resource would focus on 'range and flexibility', and 'accuracy'; Vocabulary Resource would focus on 'range' and 'appropriacy'. Apart from the use of 'range' for both Grammar and Vocabulary, it was felt that no other focus should appear in more than one scale.

2. Band descriptors for the analytical scales. The assessment criteria for another Cambridge ESOL examination, Business English Certificate (BEC) speaking are presented as a series of bullet points. OEs had commented on how much easier it was to operate with this layout so it was decided to adopt this format for the revised CPE criteria.

3. Band descriptors for the Global Achievement scale. The initial draft of the band descriptors for the analytical scales was considered and modified five times by the team and comments gathered from the STLs at their annual conference in October 1999 were also fed into the process. The fifth draft was then used as the basis for producing the first draft of the band descriptors for the Global Achievement scale. At this point it was decided to trial the assessment criteria.

PHASE 2: ASSESSMENT EXERCISE 1 (JULY 2000)

The purpose of the assessment exercise was to identify what elements of the Revised CPE analytical scales were interpreted differently by OEs, and was carried out with the following questions in mind:

- To what extent do raters agree in their assessments of a given candidate on each scale?
- What difficulties do raters face when applying the criteria?

Eighteen candidates were assessed from video footage, using Revised CPE Speaking Test materials. The candidates were of mixed ability and were of the following nationalities: Brazilian, Bulgarian, Dutch, Finnish, French, German, Greek, Swedish, Swiss German and Swiss Italian. Thirteen raters took part in the exercise: nine raters used the five analytical scales; four raters used the Global Achievement scale. The raters selected for the assessment study were all highly experienced UK OEs or Upper Main Suite TLs. These OEs attend an annual co-ordination session where their assessing capabilities are monitored. They are then monitored during live examining sessions by UK STLs, and are also involved in the monitoring of OEs throughout the Speaking Test periods. They had just completed the summer 2000 examining session, which included examining and monitoring at CPE level. Therefore, as some of Cambridge ESOL's most experienced examiners, it was felt that their differences in opinion of a candidate's performance might reflect more the inadequacies of the assessment criteria than their inadequacies as raters.

Ten days prior to the assessment exercise, the raters were sent copies of the initial draft of the Explanations of Criteria, the draft assessment criteria, and materials used by the candidates on video, in order to prepare for the assessment exercise.

At the beginning of the exercise, a CPE candidate from the 2000 Standardisation Video was shown in order to remind the raters of the level. This candidate had been assessed using the 2000 assessment criteria and was judged to be on the borderline or just below the level of adequacy for all aspects of language for a CPE Grade C. It should be remembered that, unlike other speaking tests which judge a candidate's performance across the entire range of speaking ability from complete beginner to advanced user, the assessment criteria for the Main Suite Common Scale for Speaking focus on slices of language ability, in the case of CPE at level 5.

The speaking tests were then watched in real time and each rater completed a mark sheet for each pair of candidates. Three types of data were collected:

1. Marks for each scale, which were inputted into an Excel spreadsheet and the figures were then transferred to SPSS for Windows in order to produce scores awarded to candidates on the analytical scales and descriptive statistics by candidate and raters.

2. Comments relating to use of the criteria, materials and candidate performance, which were written on the mark sheet by the raters while they were applying the assessment scales.

3. Comments relating to the assessment criteria during discussion, where the raters were asked not to discuss their marks until the end of the session when a further hour was set aside for this. This discussion was recorded on mini-disk and transcribed.

Findings

To what extent do raters agree in their assessments of a given candidate on each scale?

An analysis of the mean scores for the five analytical scales showed that generally candidates achieved the highest mean scores on the Pronunciation scale (average 3.64) and the lowest mean scores on the Grammatical Resource scale (average 3.14). For the Grammatical Resource, Lexical Resource, Interactive Communication and Global Achievement scales the raters tended to differ by 1–1.5 marks for each candidate. However, for the Discourse Management scale the raters tended to differ by 1–2 marks, and with the Pronunciation Scale the raters marks tended to differ by 1.5–2 marks.

These differences of up to 2 marks being awarded by raters for any given candidate for a particular scale suggest that, among other things, raters might be interpreting the wording of the scale definitions differently. The comments which were written on the mark sheets and the discussion which followed the exercise supported this theory.

What difficulties do raters face when applying the criteria?

A number of questions arose from the observations made by the raters and these were dealt with in Phase 3 of the development, which involved Subject Officers for the Revised CPE Speaking Test, the Assessment Criteria Developer, and the Chairs of FCE and CAE item writing teams.

PHASE 3: MODIFICATIONS TO DRAFT ANALYTICAL SCALES AND EXPLANATIONS OF CRITERIA (AUGUST 2000)

A meeting was arranged to discuss the findings of Assessment Exercise 1 and to make appropriate amendments to the Explanations of Criteria and the Analytical Scales. In preparation, the team members were sent copies of the criteria and explanations together with the statistical data, raters' notes, and the transcription from the assessment exercise. At this point, a number of changes were made, including:

- reducing the number of foci for each scale and renaming certain foci to make them examiner friendly;
- allowing a minimal degree of error at the top of the scale to ensure OEs would feel able to award the full range of marks;
- paying attention to the wording of the descriptors to make as great as possible the differentiation between 1.0 and 3.0, and between 3.0 and 5.0.

At this point, it was decided not to pursue further development of the Global Achievement Scale until the Analytical Scales had been finalised.

PHASE 4: ASSESSMENT EXERCISE 2 (SEPTEMBER 2000)

It was decided to trial the draft Assessment Criteria using Cambridge ESOL STLs at their annual conference. As for Assessment Exercise 1, the purpose of the exercise was to identify what elements of the Revised CPE analytical scales were interpreted differently by raters. The exercise was carried out with the following questions in mind:

- To what extent do raters agree in their overall assessments of the candidates?
- Which scale(s) do raters find difficult to agree upon?

Six candidates were assessed from video footage. The candidates were of mixed ability and were of the following nationalities: Bulgarian, French, German, and Greek. 26 STLs took part in the trial and these were divided into two groups.

Information provided by Cambridge ESOL Performance Testing Unit showed that of the 26 STLs, 17 had had experience of examining at CPE level within the previous 18 months (Group A); 9 had no experience of examining at CPE (Group B). Four of the STLs from Group B had not been trained to assess CPE Speaking but two of them were experienced examiners for BEC 3, which is placed at ALTE Level 4.

At the beginning of the exercise, the STLs were given an overview of the study and were then given 20 minutes to familiarise themselves with the Assessment Scales and the Explanations of Criteria and to clarify issues with the Subject Officers.

Three extracts from Revised CPE Speaking Tests were watched in real time. These extracts comprised:

- the collaborative phase of Part 2;
- long turns by each candidate;
- response and follow-up questions for each candidate.

Although the extracts did not make up a complete test, it was felt that this selection showed candidates:

- discussing something together and working towards a joint decision;
- sustaining an extended piece of discourse;
- responding to questions posed by the interlocutor and developing the topic of the discussion in depth.

Raters awarded marks to each candidate according to the revised criteria. They also wrote comments relating to their application of the criteria onto the mark sheet.

Findings

The results were analysed using Multi Faceted Rasch Analysis (Linacre 1993). This allows one to explore a number of different facets of a particular study at the same time. For this study, the purpose was to explore the data from the perspective of rater harshness and the raters' application of the scale criteria. In addition to this, a bias interaction analysis was performed. Wigglesworth (1993) suggests that in this type of analysis z-scores (a standard score which is expressed in units of standard deviation) greater than +2 and less than –2 indicate that there is significant bias. For the purposes of the study, it was suggested that bias might indicate where raters were experiencing difficulty in applying the scales.

The number of candidates involved in this study was relatively small. However, the selection of candidates representing a range of abilities and nationalities, and the use of multi-faceted Rasch analysis gave weight to the design. Also, the analysis was of value because a larger group of 26 raters was used, providing a sufficiently large group of data points. In the dataset created for this study, there are 6 candidates. Each candidate is awarded 5 scores by each of 26 raters, giving a total of 780 responses, and this is considered to be sufficient (see Wigglesworth 1993, McNamara 1996, O'Sullivan 2000).

The analysis suggested that overall there was relatively little disagreement among the STLs, with the exception of three STLs (2,7,21) who deviated by more than 1 logit from the standard in their overall assessment of the candidates (Table 1). However, all three were from Group B, i.e. they had no experience of examining at CPE level. Furthermore, they were all stationed in the Far East and may have had limited contact with European candidates as shown on the video.

However, FACETS is also able to indicate the levels of consistency displayed by the raters, whether or not the raters are applying the scales appropriately, and the relative differences between the raters in terms of their assessments. The analysis showed that four of the raters were less consistent than the other raters, that one rater had a tendency towards middle scale category overuse, and that there were significant differences between the raters in terms of harshness.

It was also observed that, as in the previous study, the criterion Interactive Communication was scored leniently, although Pronunciation was scored most harshly. The analysis suggested that these two criteria were also applied slightly less consistently than the other three criteria.

In addition to the main analysis, a bias analysis was performed in order to investigate any rater by scale interaction. There was only one instance of significant bias, and a second less significant instance of bias on the Discourse Management scale.

The range of z-scores for each of the criteria (Table 2) highlights Discourse Management as being different from the other criteria and therefore in need of attention.

Table 1: All Facet Vertical Summary (Assessment exercise 2)

```
Model =?B,?,?B,R9
---------------------------------------------------------------------------
|Measr|-Rater                                    |-Candidates|+Scale|S.1  |
---------------------------------------------------------------------------
+   3 +                                          +           +      +(9)  +
|     |                                          |           |      |     |
|     |                                          |           |      | 8   |
|     |                                          |           |      |     |
|     |                                          |           |      |     |
|     |                                          |           |      |     |
|     |                                          |           |      |     |
|     | STL7                                     |           |      | --- |
|     |                                          |           |      |     |
+   2 +                                          +           +      +     +
|     |                                          |           |      |     |
|     |                                          | C1        |      |     |
|     |                                          |           |      | 7   |
|     | STL21                                    |           |      |     |
|     |                                          |           | IC   |     |
|     |                                          |           | DM   |     |
:     :                                          :           : LR   :     :
|     |                                          |           | GR   | --- |
|     |                                          |           | Pron |     |
|     |                                          |           |      |     |
+   1 +                                          +           +      +     +
|     |                                          |           |      | 6   |
|     |                                          |           |      |     |
|     |                                          |           |      |     |
|     |                                          |           |      |     |
|     | STL3   STL4                              |           |      | --- |
|     |                                          |           |      |     |
|     |                                          | C2        |      |     |
|     | STL10  STL11  STL14  STL23  STL25  STL26 |           |      |     |
:     : STL6                                     :           :      :     :
|     |                                          |           |      | 5   |
*   0 *                                          * C3  C4    *      *     *
|     | STL15  STL16  STL20                      |           |      |     |
|     | STL22  STL24  STL9                       |           |      |     |
|     | STL12  STL17                             |           |      | --- |
|     | STL1   STL13                             |           |      |     |
|     |                                          | C6        |      |     |
|     | STL5                                     |           |      |     |
|     | STL18                                    |           |      |     |
|     | STL19                                    |           |      | 4   |
|     | STL8                                     |           |      |     |
+  -1 +                                          +           +      +     +
|     | STL2                                     |           |      |     |
|     |                                          |           |      |     |
|     |                                          |           |      | --- |
|     |                                          |           |      |     |
|     |                                          |           |      |     |
|     |                                          | C5        |      |     |
|     |                                          |           |      |     |
|     |                                          |           |      | 3   |
|     |                                          |           |      |     |
+  -2 +                                          +           +      +(1)  +
---------------------------------------------------------------------------
|Measr|-Rater                                    |-Candidates|+Scale|S.1  |
-------------------------------------------------------------------------------
```

Table 2: Range of z-scores (Assessment exercise 2)

					Range
Pronunciation	1.7	→	–0.9	=	2.6
Grammatical Resource	1.4	→	–1.1	=	2.5
Lexical Resource	1.4	→	–1.2	=	2.6
Discourse Management	1.7	→	–2.5	=	4.2
Interactive Communication	1.5	→	–1.3	=	2.8

PHASE 5: MODIFICATIONS TO ANALYTICAL SCALES AND EXPLANATIONS OF CRITERIA (SEPTEMBER 2000)

An email discussion among the group took place over a period of two weeks, in which issues raised by the STLs relating to each of the Analytical Scales were circulated.

It was felt that in some cases the issues impacted on the wording of the band descriptors while others could be addressed through examiner training. For example, under Grammatical Resource, candidates who fulfilled the *with ease and flexibility* criteria well at the expense of *accuracy* would automatically lose marks and go down to a 4.0 or 4.5. Both aspects would have to be evidenced for a mark of 5.0. This was felt to be an examiner training issue. Likewise, in terms of Lexical Resource, candidates would need both appropriate lexis and the ability to use it with flexibility to be awarded a 5.0, but would lose marks if they had one without the other. It was felt that however sophisticated lexis may seem, it is not worth crediting if used inappropriately.

One concern voiced by the STLs referred to there being a possible overlap between the Discourse Management and Interactive Communication scales. The difference between the two scales is that, one deals with the contributions being made by an individual candidate (DM) while the other deals with the skill of listening to someone else's contribution and responding appropriately (IC). It was felt that this issue should also be dealt with through examiner training.

It was suggested that in band 3.0 of Discourse Management the adverb *generally* should be introduced before *relevant.* However, although this suggestion would create a greater distance between the descriptors for bands 3.0 and 5.0 since the first words of both bands were the same, it was felt that at CPE level candidates' contributions should be relevant to be deemed adequate. The wording of the descriptors which focused on the difference between *adequate* and *fully and effectively*, and between *usually* and *consistently*, in conjunction with exemplification on standardisation videos, should address this issue. Also, in band 3.0 of Discourse Management it was agreed that, for the sake of consistency, it would be better to replace *suitable* with *appropriate,* i.e. *contributions are usually of an appropriate length.*

The Pronunciation scale generated the most discussion since it was felt that the descriptors for bands 3.0 and 5.0 were very similar and did not help OEs to discriminate the additional three bands in between (3.5, 4.0, 4.5). It had been observed that, in relation to the Common Scale for Speaking, pronunciation does not seem to improve much after CAE. However, a firm distinction had to be made for the purposes of awarding marks at CPE level. At this point, it was decided to include the aspect of 'strain on the listener' into band 3.0 to widen the gap between bands 3.0 and 5.0. However, when the criteria for CPE were later viewed in relation to the scales for the other levels in the Common Scale for Speaking, it was decided to introduce 'strain on the listener' only in band 1.0.

It was also suggested that the term *L1* be replaced with *foreign accent.* Although it was agreed that it was not always possible to trace a candidate's accent to their L1, the term has been consistently used in the assessment criteria for Cambridge ESOL speaking tests. Referring to an *accent* might introduce confusion with regional variations and the term *foreign accent* was not felt to be significantly better than *L1.* However, the issue was part of a larger, international debate and it was decided to seek further opinion.

General comments picked up on the use of *may* rather than adverbs such as *often* or *sometimes*, particularly in the descriptors for band 1.0. The rationale for using 'may' is that only one of the aspects in a particular band descriptor may be true, while the use of *often* or *sometimes* automatically assumes there is, for example, a lack of turn taking. It was felt that the use of 'and/or' addressed this issue sufficiently well.

It was agreed to rearrange the order of the bullet points to address the wider aspects of language before focusing on the specific detail. For example: *range* of grammatical structure was presented before *accuracy*; the *range* of lexis before the *appropriate* or *precise* use; *stress, rhythm and intonation* before *individual sounds.* Also, it was decided that the Explanations document should specify only the elements of language being assessed, with no reference to level.

The Explanation of Criteria and Analytical Scales were then redrafted taking account of all of these issues.

PHASE 6 COMPARABILITY OF LEVEL STUDY (OCTOBER 2000)

From the outset of the CPE Revision project it had been agreed that the level of difficulty of the examination should not change, neither for the examination as a whole, nor for specific papers. Therefore, in order to establish how closely the Revised Assessment Scales matched the previous scales, a further study was set up.

Sixteen candidates were assessed from video footage, using Revised CPE Speaking Test materials. The candidates were of mixed ability and were of the following nationalities: Argentinean, Belgian, German, Italian, Polish, Swedish, Swiss French, Swiss German, Swiss Italian, Taiwanese, Turkish. Two of the candidates had been assessed in the Phase 2 exercise, the remaining 14 candidates had not been previously assessed by the examiners.

Eight raters were selected from the Phase 2 group because they were already familiar with the format of the Revised Assessment Scales. One of the raters had acted as interlocutor for some of the tests from the new footage, but none had been involved in the continuing discussions and developments of the assessment criteria.

The raters were divided into two groups, A and B. Each group was sent:

- the previous CPE assessment scales;
- the revised CPE assessment scales;
- Explanations of the Criteria;
- separate mark sheets for the previous and revised assessment scales;
- two video tapes, each containing 5 Speaking Tests.

The video tapes were arranged as follows:

Tape 1

Test	**1**	**2**	**3**	**4**	**5**
Candidates	1A & 1B	2A & 2B	3A & 3B	4A & 4B	1A & 1B

Tape 2

Test	**6**	**7**	**8**	**9**	**10**
Candidates	5A & 5B	6A & 6B	7A & 7B	8A & 8B	5A & 5B

Each group was asked to follow a specific procedure in order to:

- ensure standardisation;
- obtain marks awarded on candidate performance using the old and the revised assessment scales in different ways;
- each rater assessed candidates 1A, 1B, 5A and 5B using both the previous and the revised scales;
- raters in Group A marked half the remaining tests using the previous scales and the other half of the remaining tests using the revised scales;
- raters in Group B mirrored the exercise carried out by Group A.

The groups were instructed as follows:

Day 1:

Remind yourself of the previous CPE 5 assessment criteria (10 minutes is allowed for this)

View tests (1-4: Group A) (6-9: Group B) and award marks for each of the candidates using the previous scales

Familiarise yourself with the revised CPE 5 assessment criteria (20 minutes is allowed for this)

View test (5: Group A) (10: Group B) and award marks for each of the candidates using the revised scales

Day 2:

Remind yourself of the revised CPE 5 assessment criteria (10 minutes is allowed for this)

View tests (6-9: Group A) (1-4: Group B) and award marks for each of the candidates using the revised scales

Familiarise yourself with the previous CPE 5 assessment criteria (5 minutes is allowed for this)

View test (10: Group A) (5: Group B) and award marks for each of the candidates using the previous scales

Marks were awarded for each scale and these were organised into two groups:

- marks awarded to the candidates who had been assessed using both sets of criteria;
- the complete dataset.

Findings

In order to gain an impression of how the scales compared, the revised marks were displayed alongside those awarded for the previous scales, inputted into an Excel spreadsheet and displayed using the Pivot Tables function to identify raters' scores for each scale.

The marks available for each scale on both the previous and the revised CPE assessment criteria range from 1 to 5. However, the previous criteria offer only whole numbers whereas the revised criteria also offer .5 of a mark.

With both the previous and revised assessment criteria the perceived level of adequate performance is based on an aggregate score of all the analytical scales. In the case of the previous scales this equates to 21/30 or 70% of the total marks. With the revised scales a mark of 15/25 or 60% is considered to demonstrate an adequate level of performance.

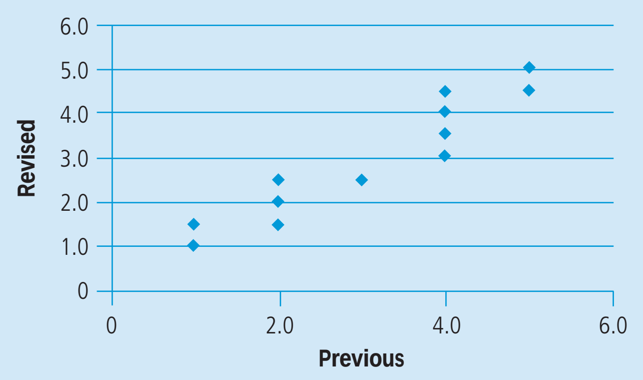
With the previous criteria the flattest profile a candidate could receive in order to receive the adequacy mark of 21 would be a combination of 3.0s and 4.0s. It is not possible for examiners to award a mark of 3.5 for an individual scale. Candidates who scored 3.0 for each analytical scale would be regarded as inadequate in their performance. Therefore we can say that 3.0 is inadequate and 4.0 is adequate. With the revised criteria a flat profile of 3.0 for each of the analytical scales is considered adequate. Therefore we can say 2.5 is inadequate and 3.0 is adequate.

The comparison of marks awarded on the previous and revised scales is affected by two factors:

1. The concept of each mark should be seen as a band ranging from the midway points either side of the mark, i.e. mark 3.0 ranges from 2.5 to 3.5. However, the scale only ranges from 1 to 5 (there are no marks of 0.5 or 5.5) so the perception of a mark of 1.0 can only range from 1.0 to 1.5, and for 5.0 from 4.5 to 5.0.
2. There is only one point where the marks are fixed: 3.0 (previous) is fixed to 2.5 (revised).

Table 3 shows how the marks are fixed.

Table 3: Comparison of marks



A rater who awards a candidate a 3.0 using the previous scales should award a mark of 2.5 using the revised scales. However, there is a range of possible revised marks which correspond to the previous marks of 1.0, 2.0, 4.0 and 5.0.

Scales on each set of criteria were matched as follows:

Previous				**Revised**
Grammatical Accuracy	GA	=	GR	Grammatical Resource
Vocabulary Resource	VR	=	LR	Lexical Resource
Interactive Communication	InC	=	IC	Interactive Communication
Pron: Prosodic Features	PP	=	P	Pronunciation
Pron: Individual Sounds	PI	=	P	Pronunciation

Both elements of Pronunciation on the previous scales were considered in relation to the element of Pronunciation on the revised scales. Fluency (previous) and Discourse Management (revised) were not thought to be similar enough for comparison.

The results of the comparability exercise showed that there was, not surprisingly, considerable agreement with marks at the upper and lower ends of the scale where the range of marks for comparison is wide: 96/105 (91%). At the adequate/inadequate boundary, with only one mark available for comparison, there were 32/55 (58%) instances where raters considered a candidate's performance to be inadequate using the previous scales but adequate when using the revised scales. However, this seemed to be a reasonable split since there is no .5 of a mark for the previous scales.

As for the STL exercise, the results from the TLs using the revised criteria were analysed using Multi Faceted Rasch Analysis. Again, the purpose of this was to explore the data from the perspective of rater harshness and the raters' application of the scale criteria. In addition to this, a bias interaction analysis was performed.

The bias analysis showed that five of the eight raters were in agreement when awarding marks using the revised scales. Three of the TLs were seen to be slightly less consistent than the other TLs, and the range of marks (Fair Average) awarded by the TLs was greater than in the STL exercise (from 3.8 to 6.0, i.e. 2.2 or 2.24 logits).

The Scale Measurement Report indicated that, with the exception of Pronunciation, the raters awarded marks with the same pattern of leniency/harshness as in previous exercises and historical trends, i.e. Interactive Communication was marked most leniently, followed by Discourse Management, Lexical Resource and finally Grammatical Resource. Anecdotal evidence had suggested that Pronunciation was different from the other criteria, and the three exercises in this study seemed to be suggesting the same. In this exercise, Pronunciation was placed in the middle of the five criteria; in the STL exercise it had been the most severely marked criterion; in the first TL exercise, however, Pronunciation had been marked most leniently. It was felt appropriate to investigate this in further studies.

The z-score was again used to identify raters who showed significant bias on a particular scale and so highlight which scales were proving problematic to the raters. There was only one instance of bias (TL5 – Pronunciation). However, a comparison of the range of z-scores (Table 4) highlighted that Pronunciation was behaving differently from the other criteria.

Table 4: Range of z-scores (Comparability of Level Study)

					Range
Pronunciation	2.5	→	–1.1	=	3.6
Grammatical Resource	0.9	→	–0.9	=	1.8
Lexical Resource	1.4	→	–0.8	=	2.2
Discourse Management	0.5	→	–1.3	=	1.8
Interactive Communication	0.7	→	–1.6	=	2.3

At this point, the Revised CPE Speaking Test Assessment Criteria were approved by the development team (with the proviso that Pronunciation would continue to be monitored) and carried forward to the final stage of development.

PHASE 7: IMPACT ON PET, FCE AND CAE ASSESSMENT CRITERIA (NOVEMBER 2000)

A working party was set up to investigate how the Revised CPE Assessment Criteria might impact on the assessment criteria for PET, FCE and CAE and what modifications might be necessary to further harmonize the Common Scale of Assessment for Speaking.

It was agreed that a generic Explanations document covering all levels was appropriate for a Common Scale of Assessment and that it should contain not only the aspects of language contained within each scale but also a summary of the overall focus of the scale. The only difference between CPE and the other scales would be the separation of Grammar and Lexis.

Further minor adjustments were made to the revised CPE scales in order to make a clear separation between the CAE and CPE scales and, after considerable debate, it was decided to take out all references to L1. However, no changes were made to the quality or level of language expected at each mark band.

A further study was carried out looking at the combined data of candidate ability across the three levels of PET, FCE and CAE to check the comparability of level with the revised and previous scales. This reinforced the view that the new format scales were examiner friendly (observations from examiners) and worked well.

At this point (December 2000), the Explanations of Criteria and Assessment Scales were approved by the Management Team with the proviso that they would be reviewed after analysis of the findings from the Standardisation Video assessment exercise. Table 5 shows the complete Explanations of Criteria.

PHASE 8: STANDARDISATION VIDEO ASSESSMENT (FEBRUARY 2001)

The making of the Standardisation Video for the CPE Speaking Test was part of an exercise which included all the Main Suite speaking tests.

Cambridge ESOL OEs are required to attend a 'co-ordination' meeting every 12 months, just prior to the main examining period, in order to re-establish the standard for each of the Cambridge ESOL speaking tests. For this purpose, Standardisation videos are produced every two years for each of the speaking tests. These show a range of candidate performances and take account of different levels of ability, different nationalities and first languages, and different combinations of male/female pairings.

Once a representative selection has been made, the performances are assessed by those STLs who have had most experience of examining at the different levels. Typically, 12 STLs who represent the major languages are involved at this stage. The STLs assess the performances in 'real' time, following a set of procedures which take account of the historical standard, and then send their independent marks to Cambridge ESOL where the marks are collated and analysed. The mark that is allocated to each candidate for each scale is based on the mean score from the STL exercise.

These marks are then sent to experienced UK TLs who scrutinise the performances for evidence to support the agreed STL mark and then write a commentary to justify that mark. The commentaries are finally examined by a team comprising highly experienced UK examiners and Cambridge ESOL Subject Officers concerned with the speaking tests.

Of the 33 candidates who were filmed for the 2002/2003 Standardisation Video, 24 were selected for inclusion in the assessment process. Eighteen candidates took the test in pairs and six took the test in groups of three, making a total of 11 tests. The candidates were selected:

- to reflect a variety of nationalities and L1;
- to show a range of abilities within the CPE level.

Twenty raters at STL and TL level who had the most, recent experience of examining and monitoring at the different levels were selected to assess the candidates on video. Twelve raters assessed the Lower Main Suite (LMS) tests of KET and PET and, in order to provide continuity across the levels, 3 of these raters also assessed the Upper Main Suite (UMS) tests of FCE, CAE and CPE, together with 8 different raters.

Table 5: Explanations of Criteria

Assessment
Candidates are assessed on their own individual performance and not in relation to each other, according to the following five analytical criteria: Grammatical Resource, Vocabulary Resource, Discourse Management, Pronunciation and Interactive Communication. These criteria are interpreted at CPE level. Assessment is based on performance in the whole test and is not related to particular parts of the test.

Both examiners assess the candidates. The Assessor applies detailed, analytical scales, and the Interlocutor applies the Global Achievement Scale, which is based on the analytical scales.

Grammatical Resource
This refers to the accurate application of grammatical rules and the effective arrangement of words in utterances. At CPE level a wide range of grammatical forms should be used appropriately and competently. Performance is viewed in terms of the overall effectiveness of the language used.

Vocabulary Resource
This refers to the candidate's ability to use a wide and appropriate range of vocabulary to meet task requirements. At CPE level the tasks require candidates to express precise meanings, attitudes and opinions and to be able to convey abstract ideas. Although candidates may lack specialised vocabulary when dealing with unfamiliar topics, it should not in general terms be necessary to resort to simplification. Performance is viewed in terms of the overall effectiveness of the language used.

Discourse Management
This refers to the candidate's ability to link utterances together to form coherent monologue and contributions to dialogue. The utterances should be relevant to the tasks and to preceding utterances in the discourse. The discourse produced should be at a level of complexity appropriate to CPE level and the utterances should be arranged logically to develop the themes or arguments required by the tasks. The extent of contributions should be appropriate, i.e. long or short as required at a particular point in the dynamic development of the discourse in order to achieve the task.

Pronunciation
This refers to the candidate's ability to produce easily comprehensible utterances to fulfil the task requirements. At CPE level, acceptable pronunciation should be achieved by the appropriate use of strong and weak syllables, the smooth linking of words and the effective highlighting of information-bearing words. Intonation, which includes the use of a sufficiently wide pitch range, should be used effectively to convey meaning, and articulation of individual sounds should be sufficiently clear for words to be easily understood. Examiners put themselves in the position of the non-EFL specialist and assess the overall impact of the communication and the degree of effort required to understand the candidate.

Interactive Communication
This refers to the candidate's ability to take an active part in the development of the discourse, showing sensitivity to turn taking and without undue hesitation. It requires the ability to participate competently in the range of interactive situations in the test and to develop discussions on a range of topics by initiating and responding appropriately. It also refers to the deployment of strategies to maintain and repair interaction at an appropriate level throughout the test so that the tasks can be fulfilled.

Global Achievement Scale
This refers to the candidate's overall performance throughout the test.

CPE Typical Minimum Adequate Performance
Develops the interaction with contributions which are relevant, coherent, and of an appropriate length. The range of grammatical forms and vocabulary is appropriate and used with sufficient accuracy and precision to deal with the CPE level tasks. Utterances are conveyed effectively and understood with very little strain on the listener.

The raters were instructed to assess the tests in order. In other words, the 11 raters who assessed the UMS Speaking Tests started with FCE, moved on to CAE and finished with CPE. This was intended to give the raters a sense of the progressive abilities of the candidates and to assess each level in relation to the adjacent levels. The assessments for FCE, CAE and CPE were carried out using the revised assessment scales.

Findings

As for the STL and TL Comparison exercises, the results from the Standardisation Video Assessment were analysed using Multi Faceted Rasch Analysis. Again, the purpose of this was to explore the data from the perspective of rater harshness and the raters' application of the scale criteria. In the dataset created for this study, there are 24 candidates. Each candidate is awarded 5 scores by each of 11 raters, giving a total of 1320 responses.

The analysis indicated very little disagreement among the STLs, an improvement on the first STL exercise, and a more consistent application of the scales. As observed in the previous studies, the raters awarded marks with the same pattern of leniency/harshness as in previous exercises and historical trends, i.e. Interactive Communication was marked most leniently, followed by Discourse Management, Lexical Resource and finally Grammatical Resource. In this exercise, unlike the previous exercise involving the STLs, Pronunciation was assessed quite leniently.

In addition to the main analysis, the bias analysis which was performed in order to investigate any rater by scale interaction indicated that the effect of linking the CPE scales into the Common Scale for Speaking across five levels may have had an effect on the raters. As a result, it was decided to take into account comments on the raters' application of the scales at the final stage of the development (Phase 9). Here is an extract from the Common Scale for Speaking.

CPE – CAMBRIDGE LEVEL 5

Fully operational command of the spoken language

- able to handle communication in most situations, including unfamiliar or unexpected ones
- able to use accurate and appropriate linguistic resources to express complex ideas and concepts, and produce extended discourse that is coherent and always easy to follow
- rarely produces inaccuracies and inappropriacies
- pronunciation is easily understood and prosodic features are used effectively; many features, including pausing and hesitation, are 'native-like'

PHASE 9: GLOBAL ACHIEVEMENT SCALE (MAY 2001)

Having completed the analytical scales, the development team returned to the Global Achievement Scale. It was felt that the initial draft was too wordy and that a more concise document should be produced which could be processed more effectively in live examination conditions. Also, as with the analytical scales, the revised CPE Global Achievement Scales should be developed alongside those for PET, FCE and CAE.

Initial drafts were drawn up and the team was joined by the Chair and Subject Officer for the PET speaking test to discuss these drafts. This meeting also provided the opportunity to make minor adjustments to the analytical scales in light of the analysis of the data from the Standardisation Video marking exercise. The amended Global Achievement Scale was then used to allocate Global Achievement scores for each of the candidates on the CPE Standardisation Video.

Conclusion

The resulting set of Assessment Scales for the Main Suite speaking tests of FCE, CAE and CPE were introduced for the December 2002 administration. When asked, Oral Examiners who were trained in their use prior to this first administration said they felt confident about being able to apply the scales. This view was repeated after the first and second administrations. Also, analysis of live examination data showed that the average mean score for each of the speaking tests was in line with historical norms, suggesting that the overall application of the Assessment Scales was consistent with previous models. Cambridge ESOL reserves the right to withhold the entire Assessment Scales.

Further information relating to this study can be found in: Weir C & Milanovic M (eds). *Studies in Language Testing 15,* Continuity and Innovation: Revising the Cambridge Proficiency in English Examination 1913–2002. Cambridge: Cambridge University Press.

References and Further Reading

Linacre, J M (1993): *FACETS Version No. 2.70* Copyright © 1987–1993.

McNamara, T F (1996): *Measuring second language performance: a new era in language testing,* London: Longman.

O' Sullivan, B (2000): *Towards a model of Performance in Oral Language Testing,* Unpublished PhD thesis: University of Reading.

Taylor, L (2000): Stakeholders in language testing, *Research Notes 2,* 2–4.

Wigglesworth, G (1993): Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction, *Language Testing* 10/3, 305–335.

CELS Speaking Assessment: towards an understanding of oral examiner and test-taker behaviour

STUART D SHAW, RESEARCH AND VALIDATION GROUP

Introduction

This article focuses on the results of an analysis of the CELS Speaking winter 2002 administration and attempts to address and provide more comprehensive insight into issues such as:

- background characteristics of the CELS candidature;
- examiner use of assessment criteria and scales;
- proper and appropriate use of the available range of task materials;
- task impact on candidate performance;
- evidence or otherwise of increasing candidate familiarity throughout the speaking test window;
- levels of agreement between Interlocutor and Assessor ratings.

Background to the CELS Speaking Test

The CELS speaking test is a standalone test of speaking in the context of general English proficiency – in a range of different contexts including social, vocational and training – whose aim is to assess English language competence through a variety of authentic tasks. The validity of the test derives from its content and format, which require candidates to engage with tasks based on real-life topics and activities involving real-life interaction patterns. Reliability is ensured through careful task design, standardised delivery, paired examiners, criteria and scales for assessment, and training and monitoring of examiners.

The CELS Test of Speaking is offered at three levels of proficiency – Preliminary, Vantage and Higher. CELS Preliminary is a lower intermediate level representing Cambridge Level 2 (and Common European Framework [CEF] level B1). At this level, candidates have a limited but effective command of the spoken language and are able to handle communication in most familiar situations. CELS Vantage is an intermediate level representing Cambridge Level 3 (CEF B2). At this level, candidates have a generally effective command of the spoken language and are able to handle communication in familiar situations. CELS Higher conforms to Cambridge Level 4 (CEF C1). At this level, candidates are expected to have a good operational command of the spoken language and are able to handle communication in most situations.

All three levels of the Test of Speaking share features in test formats, test materials, test environments, assessment procedures and criteria, and the standardisation of Oral Examiners which are described below.

Distinguishing features of the CELS Speaking Test

The common characteristics of the CELS Test of Speaking are:

1. *The Paired Test Format* – the standard test format is 2:2, i.e. two candidates and two examiners. However, where a centre has an uneven number of candidates at an examining session, the last candidate must join the final pair of candidates to form a group of three i.e. a 2:3 test format.
2. *A Two-part Test* – at each level, the CELS Test of Speaking has two parts. The test lasts a total of 20 minutes with each part lasting 10 minutes. Each part is designed to elicit a different type of language and use a different interaction pattern and both parts are equally important for assessment purposes.
3. *The Use of an Interlocutor Frame* – an interlocutor frame is a script for the examiner's role in setting up the tasks and is used for the purpose of standardisation. By adhering to the script, examiners ensure that all candidates are treated fairly and equally, thereby eliminating the risk of the focus of a task changing or the level of the test being altered by use of language inappropriate to the level.
4. *A Choice of Test Materials* – a choice of test materials is always available in order to cater for the large and varied international candidature and to maintain the security of the test, which is held over two weeks. All the test materials act as a stimulus to elicit language from the candidates and it is essential that as wide a variety of material as possible is used by Oral Examiners in each examination session.
5. *Assessment Procedures and Criteria* – when assessing the performance of candidates, the assessment criteria and the application of the rating scales should be viewed within the context of the Cambridge ESOL Common Scale for Speaking, described in Angela ffrench's article on page 8. The spread of ability to be assessed for a particular examination can be calibrated on the Common Scale.

   Throughout the test, candidates are assessed not in relation to each other but according to the following four equally-weighted analytic assessment criteria:

   - Grammar and Vocabulary – on this scale, candidates are awarded marks for the accurate and appropriate use of grammatical structures (in addition to syntactic forms at the Higher level) and vocabulary in order to meet the task requirements.
   - Discourse Management – on this scale, examiners are looking for evidence of the candidate's ability to express ideas and opinions coherently and effectively in order to fulfil the task, through use of a suitable range of linguistic devices and extended utterances where appropriate. The tasks require candidates to construct sentences and produce utterances (extended as appropriate) in order to convey information and to express or justify opinions. The candidate's ability to maintain a coherent flow of language with an appropriate range of linguistic resources over several utterances is assessed here.
   - Pronunciation – this refers to the candidate's ability to produce comprehensible utterances to fulfil the task requirements i.e. it refers to the production of individual sounds, the appropriate linking of words, and the use of stress and intonation to convey the intended meaning. First-language accents are acceptable provided communication is not impeded.
   - Interactive Communication – this refers to the candidate's ability to take part in the interaction with the Interlocutor and the other candidate and fulfil the task requirements by responding and initiating appropriately and at the required speed and rhythm. It includes the ability to use functional language and strategies to maintain or repair interaction, e.g. conversational turn-taking, and a willingness to develop the conversation and move the task towards a conclusion. Candidates should be able to maintain the coherence of the discussion and may, if necessary, ask the Interlocutor or the other candidate for clarification.
   - In addition to the analytic scale, a separate scale – the Global Achievement Scale – is used by the Interlocutor to assess the candidate's overall effectiveness in tackling the tasks:
     - Global Achievement Scale – the Interlocutor is required to give one global mark on the scale to reflect the candidate's performance across both parts of the test and this is not necessarily the average of the analytical marks, but an impression mark reflecting an assessment made from a different perspective.

Table 1: Test format and timing

Part	Time	Interaction pattern	Input
1	10 mins	*Preparation time – 1 min 30* Candidates talk individually with Interlocutor on prompts they have chosen	ι Written stimulus from task ι Oral stimulus from Interlocutor
2a 2b	10 mins	*Preparation time – 1 min 30* Interlocutor sets up task; candidates talk together Three-way discussion between Interlocutor and candidates	ι Written stimulus from task ι Written prompts on task ι Oral prompts from Interlocutor ι Two-way discussion from 2a

– Both the global and four analytic scales comprise 6 Band Levels (0 – 5 in half-bands). Detailed descriptors are provided for Bands 1, 3 and 5. Band 3 epitomises the standard required for a satisfactory performance at a particular level of CELS. Final ratings should reflect variations in performance across both parts of the test.

6. *A Standardised Test* – with an operation of an international scale, it is obviously crucial to ensure the standardisation of the conduct and assessment of the Test of Speaking. In order to ensure these objectives, Cambridge ESOL has set up a global framework of Oral Examiner Team Leader Systems, which began with the 17 largest countries (over 95% of the candidature) and is now spreading to cover the remaining smaller countries.
7. *Interlocutor/Assessor Examiner Assessment* – assessment is made by paired examiners, one acting as the Assessor, the other as the Interlocutor. The *Assessor* listens but takes no active role in the interaction and assesses the candidates by applying the detailed Analytical Scale. The *Interlocutor* is responsible for managing the interaction by adhering to the Interlocutor Frame and instructions, ensuring that both candidates are treated fairly and equally and observing the prescribed test timings carefully. The Interlocutor also assesses the candidates by applying the Global Scale i.e. the less detailed scale based on the Analytical Scale.
8. Each examiner views the performances from a different perspective and arrives at marks using a different set of criteria. Accordingly, examiners are not expected to discuss their marks or change their assessments in the light of those made by a co-examiner.

Oral Examiner and Candidate Performance

An understanding of how CELS participants performed during the winter 2002 speaking session enables us to address a number of validation questions related to candidature performance, oral examiner behaviour and nature of the rating scale.

1. What are the background characteristics of the CELS candidature?

Candidate Information Sheets (CIS) are routinely administered to all ESOL candidates enabling Cambridge ESOL to gather a large amount of demographic data such as age, gender, nationality, first language etc. for research purposes. CIS data reveals that:

- almost half of the candidature were Preliminary Level candidates and marginally more than one-third were Vantage Level. The Higher Level constituted the smallest proportion of candidates with slightly less than one-fifth of the total candidature;
- in terms of first language, the candidature was overwhelmingly Spanish or Portuguese;
- most candidates taking the tests were based in Uruguay, Brazil, Argentina and the UK.

2. How did examiners use the assessment criteria and scales?

The three main descriptive statistics used to describe the distribution of marks given by a set of raters are the mean, or average mark, the standard deviation, or the average amount that marks differ from the mean, and the range, which is the easiest way to talk about the spread of marks from the central mark. The important point about standard deviation is that the larger the index of standard deviation, the wider the range of distribution away from the measure of central tendency. The smaller the standard deviation index, the more similar the scores, and the more tightly clustered the data are around the mean. This has implications for the extent of the mark scheme scale employed by the examiner. Descriptive statistics reveal that:

- across all three language proficiency levels the assessment criterion *Interactive Communication* demonstrated the highest mean score and the assessment criterion *Grammar and Vocabulary* the lowest mean score;
- in general, examiners used the full range of the available scale (1–5). *Interactive Communication* and *Pronunciation* tended to use a slightly narrower range at the Preliminary Level (2–5); all 5 scales operated over slightly narrower ranges at the Higher Level (2–5);
- the *Pronunciation* criterion revealed, consistently, the lowest standard deviation for each of the three levels;
- the highest correlations occurred between the *Grammar and Vocabulary/Discourse Management* and *Discourse Management/Interactive Communication* scores;
- *Pronunciation* seemed distinctively different from other analytic criteria.

3. Did examiners use an appropriate range of materials available?

Examiners are expected to use the full range of sets of material provided. The choice of task is generally made on a random basis and examiners are not supposed to select particular tasks for particular candidates, except where sensitivity needs to be exercised with candidates from specific backgrounds, e.g. refugees or asylum seekers. Analysis of the CELS speaking data reveals that:

- generally there was a good take up of all the tasks provided;
- tasks at the front of the set tended to be used more than later ones;
- only a few tasks appeared underused;
- no test packs were used significantly more than others.

4. Did test pack impact on candidate performance?

Descriptive statistics further show that:

- tasks performed in a largely similar way in terms of mean scores with only a few tasks attracting slightly higher/lower mean scores;

- CELS Higher speaking tasks seemed to demonstrate the greatest capacity to differentiate amongst candidates.

5. Did test format impact on candidate performance?

The majority of interviews were conducted as paired tests (89% for Preliminary, 83% for Vantage, 80% for Higher). On average candidates in the 3:2 format scored at least as well or even slightly better than those in the 2:2 format, i.e. candidates in a trio were not disadvantaged. This effect was observed across all three language proficiency levels.

6. Was there any evidence of familiarity increasing throughout the Speaking Test window?

The trend in the mean total score for candidates taking the test on different days is an indicator of test material security. An upwards trend may be an indication of insecure test materials. In the analysis there appeared little or no increase in mean scores in all levels as the session progressed.

7. What level of agreement was there between Interlocutor and Assessor marks?

One way of addressing this issue is to estimate the correlation between the Interlocutor and the Assessor i.e. the combined analytic scores (as rated by the Assessor) with the global score (as rated by the Interlocutor). Correlational analyses can be used to investigate the relatedness of the two rating approaches. The magnitude of the correlation coefficient indicates how well the two sets of measurements agree. The closer the value is to 1, the stronger the relationship between the two ratings. Correlation figures for '*All Countries*' were respectable and not excessively high (.81 for Preliminary, .76 for Vantage and .85 for Higher). There was therefore no evidence that oral examiners were colluding when assessing marks.

Conclusion

This study forms part of the ongoing validation programme for the CELS speaking test and its findings help to confirm the reliability and validity of the current approach. They also contribute to the wider research programme for the testing and assessment of spoken language. Constant monitoring of test performance enables us to gain a greater appreciation of the nature, role and performance of speaking test candidates and their examiners, and can contribute to improved and standardised delivery of tasks, careful and appropriate selection of speaking materials, shared understanding and consistent application of assessment criteria and an enhanced training and monitoring programme for examiners.

References and Further Reading

Lazaraton, A (2002): A Qualitative Approach to the Validation of Oral Language Tests, in Milanovic, M and Weir, C (eds): *Studies in Language Testing 14*, Cambridge: Cambridge ESOL/CUP.

Taylor, L (2000): Issues in Speaking Assessment Research, *Research Notes* 1,8–9, Cambridge: Cambridge ESOL.

Evaluating the success of the revised BEC (Business English Certificate) Speaking Tests

DAVID BOOTH, SPECIALISED EXAMINATIONS GROUP

Introduction

This article is based on a presentation given to IATEFL Brighton in April 2003. The aim of the presentation was to inform the audience about some of the ways in which Cambridge ESOL monitor the delivery of speaking tests. In particular the presentation focussed on the data collected on the speaking test marksheet and the information that could be recovered from this regarding the conduct of the test, the scoring of the test, examiner and candidate choices within the test and information relating to test fairness. The focus throughout is on the revised BEC speaking tests which were introduced in March 2002 as part of the wider revision of the BEC suite of tests.

The article first looks at the format of the revised test to give a context for the evaluation. It then looks at the data collected from the marksheet and proposes a number of research questions. It then suggests answers to the questions based on the data collected.

The revised format

The revised format of the BEC speaking tests was introduced alongside other changes in March 2002. The changes to BEC are documented in Booth 2002 and O'Sullivan forthcoming. Some of the main changes made to the BEC suite include: the introduction of BEC Preliminary at ALTE level 2 as a replacement for BEC1 and changes to the weighting of papers so that all skills contribute 25% of the overall grade. Changes to the speaking test included changes to timings to make the test longer and the introduction of a longer turn in the Preliminary and Vantage level tests.

The revised BEC speaking test has three parts; part one is an 'interview' format where the examiner asks questions to the candidates in turn. These questions come from the scripted Interlocutor Frame. In part 2 candidates talk for about a minute on a topic which they select from a task card which the examiner gives them. In the third section candidates discuss a topic together. The topic (or in the case of Preliminary a scenario) is selected by the examiner. At Preliminary level the scenario is supported by visual or written prompts.

A key element to the test is a choice of material. Examiners can choose from a number of tasks provided for each part of the test. In the *Instructions to Oral Examiners*, a handbook for examiners, and in annual training, examiners are reminded to vary their use of material. By doing this examiners enhance the security of test material by reducing the risk of candidates predicting the content of the test. Additionally there are specific materials provided for group tests where three candidates are examined together.

Candidates also exercise choice. In part 2 of the test candidates are given a choice of topic to talk about. At Preliminary they choose 1 topic from 2. At Vantage and Higher they choose 1 from 3. Different levels of support are provided at each level.

The revised BEC speaking test follows the main Cambridge ESOL model for speaking of using two examiners for each test (see Taylor 2001). The normal format is two examiners and two candidates though there is provision for three candidates to be examined in the final test in a test session where there are an odd number of candidates. Examiners also change roles during the test session. One examiner acts as Interlocutor conducting the test; the other examiner acts as Assessor. Both examiners mark the test though using different scales. The assessor has four detailed scales covering Grammar and Vocabulary, Discourse Management, Pronunciation and Interactive Communication. The Interlocutor has a Global scale which is derived from the more detailed scales. The marks are arrived at independently and examiners are instructed not to discuss the marks they award with each other.

The reading, writing and listening components of BEC are held on a designated day in the morning. Examination centres are, however, given a window period in which candidates should take the speaking test. This gives centres the flexibility to timetable tests to suit local needs.

Data collection

The speaking test marksheet is designed to be read by an optical reader. This means that information is scanned electronically into a database. Obviously the marks awarded are collected and this is linked to candidate data which allows the processing of grades and results within a short time frame. Further information is collected regarding the test which is then analysed by Validation staff. The speaking test marksheet collects the following data:

- Centre number and name;
- Candidate number;
- Examination title, code and session;
- Date of test;
- Marks awarded for: Grammar and Vocabulary, Discourse Management, Pronunciation, Interactive Communication and Global Achievement;
- Test materials used in each part;
- Assessor, Interlocutor and other candidate ID;
- Test format.

What questions can we ask?

The data collected allow us to ask a number of questions about the conduct of the speaking test. For example information on the date of the test allows us to look at the areas described below.

The date of the test

1. *Is the test window fully exploited?*
   Centres can timetable tests at any time within a specified period. Are all available days used by centres? Which day(s) are most popular?
2. *Were any tests taken outside the window?*
   Within our centre monitoring scheme, it is important for us to know if any tests are taken outside the designated window.
3. *Is there any advantage to candidates who take the test later in the window?*
   Given the instructions to Oral Examiners to rotate materials, it is unlikely that candidates who take the test later in the window have an advantage. Clear guidelines are also in place on the day to minimise the contact between candidates who have just taken the test and those who are about to.

Marks awarded

4. *Which assessment scale tends to have the highest mean mark?*
5. *Which assessment scale tends to have the lowest mean mark?*
   Do candidates tend to do less well on Grammar and Vocabulary or Pronunciation? Do some assessment scales discriminate more than others?
6. *What is the correlation between examiner marks?*
   The Assessor and Interlocutor give separate, independent scores from slightly different perspectives. However, these should correlate to an acceptable degree.

Test materials

7. *Do examiners vary their use of materials?*
8. *Are any of the tasks easier or harder?*
9. *Which tasks do candidates choose in part 2 of the test?*

Findings

The data collected from the speaking test marksheet provides valuable information to Cambridge ESOL. It allows us to be confident that the test is being conducted properly. This data collection, however, is not done in isolation. Cambridge ESOL has put in place a Team Leader System. This system, co-ordinated from Cambridge, ensures the quality of the speaking test assessment by managing the recruitment, induction, training, co-ordination, monitoring and evaluation of speaking test examiners worldwide. Any issues which may result from the analysis described above will impact directly on the training and co-ordination of examiners.

The date of the test

The majority of BEC candidates take the speaking test on the same day as the other components. The table below illustrates this for a group of candidates who sat BEC Preliminary in May 2002. The test date was the 25th of May and the speaking test window was between the 11th and the 26th of May.

Table 1: Test Window Usage BEC Preliminary 2002 May session

Test date	% of candidates (groups larger than 2.5% of the test population)	Mean scores
17.05.02	2.7	24.3
22.05.02	5.3	24.1
23.05.02	7.1	24.5
24.05.02	7.8	23.3
25.05.02	67.8	23.5

Not all the available dates in the window were used, particularly near the beginning of the window. Whilst two thirds of the candidature took the speaking test on the same day as the other components a significant number of candidates took their speaking tests on a different date. The data indicated that some tests were taken outside the window.

The evidence from mean scores also suggests that the candidates' scores do not increase later in the test session. Similar evidence is found in subsequent administrations.

Marks awarded

The table below gives the mean marks and standard deviation (SD) for a large group of BEC Vantage candidates.

Table 2: BEC Vantage 2002 Mean score for analytical rating scales

Rating Scale	Mean	SD
Grammar and Vocabulary	3.86	.76
Discourse Management	3.99	.76
Pronunciation	4.06	.67
Interactive Communication	4.09	.73

The lowest mean score is achieved on the Grammar and Vocabulary scale. This is consistent with findings for speaking tests in other Cambridge ESOL examinations. The other three scales have very similar marks. An interesting feature of Pronunciation is the low standard deviation which indicates that the Pronunciation scale discriminates less than the other three scales, probably because the pronunciation trait is less easy to scale.

Assessor/Interlocutor agreement

The level of agreement between Assessor and Interlocutor marks should be reasonably high but not too high. Correlations above .9 may indicate that examiners are discussing their marks before entering the marks on the marksheet. The average Pearson's correlation for the top twenty countries was .83. Higher/lower correlations are dealt with through the training and co-ordination provision in the Team Leader system.

Test materials

Examiners vary their use of material. In 2002 there was no evidence of any particular task being neglected. Mean scores for tasks are monitored to check if there is any significant difference between candidate scores for individual tasks.

Conclusion

The data presented above represent one of the ways Cambridge ESOL monitors the conduct of speaking tests. The focus on the BEC speaking test is in light of the recent revisions to the speaking test introduced at the beginning of 2002. Alongside this there has been a programme of training for oral examiners which has also been monitored. Results of this were circulated to Senior Team Leaders within the Team Leader System. Additionally Cambridge ESOL is consulting with centres and key stakeholders on the revised test and on the impact of the changes. Cambridge ESOL has also published support material including sample tests and a Sample Speaking Test video to support the changes to the test. Further information on this can be found on the website.

References and Further Reading

Booth, D (2002): Revising the Business English Certificates (BEC) speaking tests, *Research Notes 8*, 4–7, Cambridge: Cambridge ESOL.

O'Sullivan, B (forthcoming): *Issues in Testing Business English; The Revision of the Cambridge Business English Certificates*, Studies in Language Testing Series 17, Cambridge: UCLES/CUP.

Taylor, L (2001): The paired speaking test format: recent studies, *Research Notes 6*, 15–17, Cambridge: Cambridge ESOL.

Conference Reports

The following reports describe several conferences attended recently by Cambridge ESOL staff: IATEFL (Paul Seddon and Trish Burrow) and AAAL and TESOL (Lynda Taylor).

IATEFL 2003: Joint CALL/TEA SIG pre-conference event

This year the CALL SIG (Computer Assisted Language Learning Special Interest Group) and TEA SIG (Testing, Assessment and Evaluation Special Interest Group) held a joint pre-conference event at the University of Brighton. The purpose was to raise awareness of computer based tests (CBTs), argue the pros and cons of using them and provide the opportunity for the participants to get hands-on experience using CBTs in the afternoon workshops. The day's two plenary speakers Glenn Fulcher (University of Dundee) and Barry O'Sullivan (University of Roehampton) had been tasked with arguing for and against the use of computers in testing and a panel discussion event was time-tabled at the end of the day for a discussion of the issues raised.

The audience, mainly made up of teachers and lecturers interested in testing, computers or both, were informed of trends in CBTs: the differences between the more traditional linear CBTs and Computer Adaptive Tests (CATs), the immense resources needed to populate CAT banks and some of the security issues Computer Based Tests have had to overcome. Participants learned that an adaptive test required a bank with a minimum of 500 items to enable accurate assessment of candidates' abilities. These itembanks provide different item types and need to cover the full range of item difficulties to adequately measure candidate performance; thus only the largest testing bodies could afford to produce CATs. It was argued that itembanks for the CATs could be 'CAT vats' (over a thousand items), 'pools' or even 'lakes' and 'oceans' depending on their sizes.

Some of the issues raised included questions such as: Is the construct of a CBT the same as its pen & paper (P&P) counterpart? Can CBTs be judged to have the same construct given that the medium of delivery is different? Or should we accept that the construct is not the same and that the two forms of testing are not equivalent and that in this case, candidates should be allowed 'bias for best' – in other words allowing candidates to choose between taking a CBT or a P&P test depending on their preferences.

The notion of equivalence between computer based and pen & paper tests was raised again in the panel discussion. It was suggested that further research is needed in order to study how reading text on a computer screen and reading text on paper are cognitively processed and how examiners respond to hand-written errors compared to typed errors. It was also asserted that in some regards, CBTs were going backwards not forwards. The item types used in CBTs were limited and conformed to traditional item types rather than taking the opportunity to develop new constructs, relevant to the emerging medium, engaging the candidate with item types of greater face validity and usefulness.

The afternoon workshops enabled participants to assess their own linguistic abilities in either French, Spanish or German using CB BULATS, a computer adaptive test produced in conjunction with KoBaLT (Computer based language testing group). KoBaLT is made up from ALTE members, Cambridge ESOL, the Alliance Française, the Goethe Institut and the Universidad de Salamanca. ETS took the opportunity to inform participants of their new phone tests and *e-rater*®, ETS's automated writing scoring software.

The consensus reached was that CBTs made a valid contribution to language assessment. Candidates needed the option of choosing bias for best yet both the CBT and its P&P stablemate needed to report on the same scale. In the future CBTs would continue to grow in use but perhaps there needed to be a rethink in terms of construct, exploration of new possibilities in item types and a greater emphasis on security especially in light of the development of 'high stakes' CBTs.

IATEFL 2003: Task design for children

At the UK IATEFL conference this year, Annamaria Pinter, from the University of Warwick, gave an account of her research into task design for Young Learners (YL). This research was prompted by the growth in popularity of tasks in YL teaching.

Aims of the research project

The project aimed to investigate if children of different age groups can cope with the demands of communicative tasks and whether children at lower levels of proficiency cope with communicative tasks.

The research project

Children aged ten, studying English in Hungary were observed over a period of three weeks performing two types of tasks: 'Spot the Difference' and 'Follow the Route on a Map'. These tasks were chosen for their child appeal, and because they allowed researchers to compare performance in a two-way task with performance in a one-way task. The tasks were completely new to the cohort. The children did the tasks four times in total. The teacher built in revision phases into the programme and provided different sets of material so that the children could improve their performance. The children were filmed and then watched this and commented on their own performance.

The tasks

The Spot the Difference task consisted of two pictures of houses that looked similar, but had six significant differences between them. The children were taught to work in pairs and find the six differences without looking at their partner's picture. In order to embed referential conflicts in the task and ensure that the children were not able to complete the task by only asking one type of question repeatedly (e.g. 'Is there a frog/dog/cat? etc.), the differences were varied and included changes such as position and activity.

The Follow the Route task was presented to the children as a game. The children worked in pairs with one child describing a route and the other following the route on a separate map. They were told that a spell had been put on the forest and that there was only one safe way to get to the monster's den. In order to complete this task fully, the children had to ensure that information was successfully encoded and decoded at every step of the journey.

Findings – Task strategies

Adults were also recorded doing the same tasks as the children and comparative transcripts were collected. These transcripts were analysed, as researchers wished to learn what strategies were employed to complete the tasks. The findings show that Young Learners:

- produce less language than adults;
- use L1 more and don't use strategies such as using gestures to keep going in L2;
- negotiate meaning far less than adults;
- operate more loosely and do not persist to clear up ambiguities;
- do not seek agreement or disagreement.

Findings – Search maps

Using the transcripts, the researchers then made 'search maps' for each interview to track the order in which the children tackled the information. The researchers drew maps with numbered turns onto a copy of the visuals used in the tasks.

For the Spot the Difference task, the aim was to discover if children used i.e. a top to bottom approach to describe their pictures, or if they employed other strategies. The aim of tracking moves in the Follow the Route task was to examine the extent to which the task could be completed correctly when misunderstandings occurred.

Comparison of search maps for the Spot the Difference task showed that adults are more systematic than children, whose strategies look 'random' to the adult eye. The children did use strategies, but these differed from 'logical' approaches, i.e. children did not work their way from left to right across the picture, but first did all the words they knew in English, or talked about all the people and then all the animals.

A comparison of the search maps for the Spot the Difference task and for the Follow the Route task showed that two-way tasks make far fewer demands on Young Learners and lead to a greater degree of success. The demands on the partners are equal, so children find them easier to do. In one-way tasks, children struggle a lot more, as there is a greater risk of failure and the children are less able to repair communication when a mistake is made.

Conclusion

Children enjoy doing tasks and cope well even when the tasks are new.

- Tasks that place fewer demands on children lead to greater success and are more appropriate for young learners with lower levels of language proficiency.
- Children approach tasks differently from adults, and they bring their own version of logic to the tasks.

The above findings show that in devising tasks to assess Young Learners' speaking abilities, we need to consider the children's reasons for doing a task, their cognitive development and ways to ensure that the children can overcome potential obstacles to successful communication. With current plans to review parts of the YLE speaking tests, this research adds to the discussion of key principles in YL task design.

References and Further Reading

Carpenter, K, Fujii, N and Kataoka, H (1995): An oral interview procedure for assessing second language abilities in children, *Language Testing* 12/2, 157–181.

Ellis, R and Heimbach R (1997): Bugs and birds: Children's acquisition of second language vocabulary through interactions, *System* 25/2, 247–259.

Oliver, R (2000): Age differences in negotiation and feedback in classroom and pairwork, *Language Learning* 50/1, 119–151.

AAAL/TESOL 2003

The annual conference of the American Association of Applied Linguistics (AAAL) was held this year in Arlington, Virginia, from 22 to 25 March – within sight of the US Pentagon; the theme for this year's conference was 'The Diversity of Applied Linguistics'. Sadly, the unpredictability of world events in March 2003 discouraged many presenters and delegates from embarking on overseas trips and led to a hasty rescheduling of presentation schedules on the part of the conference organisers. Despite this, the plenaries, papers and posters which did go ahead provided ample demonstration of the diversity of domains and issues which now characterise the field of applied linguistics: from language cognition to language and ideology, from translation and interpretation to language minority education, to name but a few. The field of assessment and evaluation was well represented and within this strand Lynda Taylor gave a paper on behalf of Cambridge ESOL entitled 'Responding to diversity: issues in assessing language learners with special needs'. Her paper described Cambridge ESOL's provision for special needs candidates in its exams and reported on some recent investigative work into the writing test performance of L2 learners with dyslexia.

The AAAL conference was followed as usual by the annual conference for Teachers of English to Speakers of Other Languages (TESOL), this year held in Baltimore, Maryland, from March 25 to 29. The conference theme was 'Hearing Every Voice' and once again there was a heavy focus upon diversity within the field, this time from a pedagogic (rather than applied linguistic) perspective. As a key exhibitor at the conference, Cambridge ESOL was able to present two important sessions related to its examination products: Susan Barduhn, until recently president of IATEFL, presented a session on Cambridge ESOL's CELTA teaching award (Certificate in English Language Teaching for Adults) which offers a valuable introductory qualification for those wishing to enter the TESOL profession; Lynda Taylor and Beryl Meiron gave an information session on the International English Language Testing System (IELTS) which is growing rapidly in popularity worldwide as an international language proficiency requirement for college entry and career development. Considerable interest was shown in both these sessions, and particularly in the IELTS session due no doubt to the growing awareness of the test and its relevance in the US context.

Research Notes Offprints

Offprints of past *Research Notes* articles are now available to view or download from the Cambridge ESOL website. The offprints from issues 1–6 are listed below and those from issues 7–13 will be listed in a future issue.

Issue	Date	Title	Author/s
Issue 1	Mar-00	EFL Research at UCLES	
		The EFL Local Item Banking System	Simon Beeston
		Developing Language Learning Questionnaires (LLQs)	Nick Saville
		Issues in Speaking Assessment Research	Lynda Taylor
		The UCLES/CUP Learner Corpus	Andrew Boyle & David Booth
Issue 2	Aug-00	Stakeholders in language testing	Lynda Taylor
		Investigating the impact of international language examinations	Nick Saville
		The UCLES EFL item banking system	Simon Beeston
		Development of new item-based tests: The gapped sentences in the revised CPE Paper 3	David Booth & Nick Saville
		Background to the validation of the ALTE 'Can-do' project and the revised Common European Framework	Neil Jones
		Investigating the paired speaking test format	Lynda Taylor
		Using observation checklists to validate speaking-test tasks	Nick Saville
Issue 3	Nov-00	Principles and practice in test development: the PETS Project in China	Lynda Taylor
		The use of Rasch Partial Credit Analysis in test development	Simon Beeston
		Developing observation checklists for speaking tests	Nick Saville & Barry O'Sullivan
		BULATS: A case study comparing computer based and paper-and-pencil tests	Neil Jones
		Approaches to rating scale revision	Lynda Taylor
		New-style statements of results	Neil Jones
Issue 4	Feb-01	Reliability as one aspect of test quality	Neil Jones
		Test Development and Revision	Nick Saville
		Revising the IELTS Speaking Test	Lynda Taylor & Neil Jones
		Announcement of the winners of the IELTS MA Dissertation Award 2000	
Issue 5	Jul-01	Revising the IELTS Speaking Test: developments in test format and task design	Lynda Taylor
		The ALTE Can Do Project and the role of measurement in constructing a proficiency framework	Neil Jones
		Towards a common scale to describe L2 writing performance	Roger Hawkey
		CB BULATS: Examining the reliability of a computer based test using test-retest method	Ardeshir Geranpayeh
Issue 6	Nov-01	Issues in the assessment of second language writing	Stuart D Shaw
		Using corpora in language testing	Fiona Ball
		Revising the IELTS Speaking Test: retraining IELTS examiners worldwide	Lynda Taylor
		The IELTS Impact Study: development and implementation	Roger Hawkey
		The paired speaking test format: recent studies	Lynda Taylor
		European language testing in a global context	Marianne Hirtzel