CAMBRIDGE

Cambridge University Press
978-0-521-01330-7 - Issues in Testing Business English: The Revision of the Cambridge Business English Certificates
Barry O'Sullivan
Excerpt
More information

# 1 Introduction to the testing of language for business purposes

## A brief historical introduction

Though there have been formal tests of general proficiency around for many years – see Weir (2003a) for an interesting and informative historical perspective on the Certificate of Proficiency in English (CPE) since its introduction in 1913 – interest in language for specific purposes has a far shorter history, emerging, according to Swales (1984:11) with Barber's (1962) *Some Measurable Characteristics of Modern Scientific Prose*. This is not to say that there has been an awareness of the use of language for specific purposes only in recent times. Schröder reminds us:

> . . . when new counting house regulations were issued for the London Salhof in 1554, these stated amongst other things that young apprentices from Germany would have to spend one year with a clothmaker in the country, so that they might get a proper command of everyday English and the more specific technical terms . . . (1981:43).

Much of the early work in the area was driven by research which focused on the identification of unique instances of language use in specific contexts (Hüllen 1981a, 1981b, Johns 1980, Lackstrom, Selinker and Trimble 1973, Selinker and Douglas 1985, Swales 1971, to list but a few), the issue of authenticity in the use of materials for teaching (e.g. Carver 1983) and the central place of needs analysis in identifying the specific language needs of learners in given contexts (Alwright and Alwright 1977, Brindley 1984, Gledhill 2000, Hawkey 1978, Hutchinson and Walters 1987, Kennedy and Bolitho 1984, LCCIEB 1972, Robinson 1980, 1985, Thurstun and Candlin 1998, West 1994). As can be seen from the dates of these publications, much of the English for Specific Purposes (ESP) debate was conducted almost twenty years ago, yet many of the same questions continue to be asked today.

Hawkey (2004) outlines the changes in theories of language learning and teaching that lead to the development of a clearly defined ESP methodology, and led to an awareness of the need to establish a set of clearly rationalised testing procedures. In the case of the testing of language for



© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-01330-7 - Issues in Testing Business English: The Revision of the Cambridge Business English Certificates
Barry O'Sullivan
Excerpt
More information



business purposes, the first test to emerge was the Test of English for International Communication (TOEIC). It was developed by Educational Testing Services (ETS) in the USA and introduced in 1979. The test, originally devised for the Japanese market, was based firmly on psychometric–structuralist theory (Spolsky 1995) and represents one of the few remaining (though highly successful from a commercial perspective) examples of a multiple-choice format, standardised, international language test.

While the TOEIC looked backwards for its theoretical underpinning, other tests of business language, particularly those developed in the UK, were beginning to look to a more communicative model. Theorists on communicative competence, particularly Canale and Swain (1980), Hymes (1972) and practitioners like Munby (1978) had a profound influence on the practice of language teaching and testing. One major influence was the facilitation of a movement away from the psychometric–structuralist methodology, based on the teaching and testing of discrete aspects of language, to the psycholinguistic–sociolinguistic era, where language teaching and testing were seen from a holistic or integrated perspective. The shift in emphasis in language teaching from language *knowledge* to language *use* paved the way for a testing methodology which reflected the same ideas. Hawkey (2004) traces the historical development of the theoretical movements of this period and provides a contextualisation for the emerging interest in the teaching and later testing of ESP. With the exception of the TOEIC, the tests described in the following sections have an essentially performance-based orientation in which emphasis is placed on the contextualisation of the tasks and predicted linguistic responses within the business setting.

In the mid-1980s the move to the testing of language for business purposes in the UK began in earnest with the development by the Royal Society of Arts (RSA) of the *Certificate in English as a Foreign Language for Secretaries* (CEFLS) – which was later administered as the *Certificate in English for International Business and Trade* (CEIBT) – and a corresponding move by the London Chamber of Commerce and Industry Examinations Board (LCCIEB) and Pitman (now part of the City and Guilds Examinations Board) to create language tests with a business focus. When the RSA was subsumed into the University of Cambridge Local Examinations Syndicate (UCLES) in 1988 the RSA test was administered by UCLES, establishing its portfolio of language tests for business.

In the early 1990s two new examinations, the Business English Certificate (BEC) and Business Language Testing System (BULATS) were developed by UCLES. It is the former of these tests that forms the basis for the latter part of this book, in which the procedures used by Cambridge ESOL in the Business English Certificate (BEC) suite revision are outlined and exemplified.

During the mid- to late-1990s a number of tests of other languages for business emerged. These included JETRO (Japanese), Test de français interna-



© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-01330-7 - Issues in Testing Business English: The Revision of the Cambridge Business English Certificates
Barry O'Sullivan
Excerpt
More information



tional (TFI) from the makers of TOEIC, the Certificate in Italian for Commerce (CIC) and the tests in the BULATS series (French, German and Spanish in addition to the English version).

There is clearly a growing interest in the area of testing language for business purposes, particularly with the internationalisation of business and the need for employees to interact in more than just a single language. The move towards a 'business language' testing genre is reflected in the tests mentioned above and described in the latter part of this chapter.

## Theoretical perspectives

In the only serious attempt to date to build a theoretical rationale for the testing of language for specific purposes, Douglas (2000) argues that a theoretical framework can be built around two principal theoretical foundations. The first of these is based on the assumption that language performance varies with the context of that performance. This assumption is supported by a well established literature in the area of sociolinguistics – see for example Labov's (1963) classic study of vowel change on Martha's Vineyard – in addition to research in the areas of second language acquisition (Dickerson 1975, Ellis 1989, Schmidt 1980, Smith 1989, Tarone 1985, 1988) and language testing (Berry, 1996, 1997, Brown 1995, 1998, Brown and Lumley 1997, O'Sullivan 1995, 2000a, 2000b, 2002a, Porter 1991a, 1991b, Porter and Shen, 1991). This fits well with the growing interest in a socio-cognitive approach to language test development where performance conditions are seen to have a symbiotic relationship with the cognitive processing involved in task completion (introduced by O'Sullivan 2000a and discussed in detail by Weir 2004).

In the case of the second foundation, Douglas sees specific purpose language tests as being 'precise' in that they will have lexical, semantic, syntactic and phonological characteristics that distinguish them from the language of more 'general purpose' contexts. This aspect of Douglas's position is also supported by an ever increasing literature, most notably in the area of corpus-based studies of language in specific contexts (Beeching 1997, Biber et al 1998, Dudley-Evans and St John 1996, Gledhill 2000, Thurstun and Candlin 1998).

When it came to an actual definition of specific purpose tests, Douglas places these two foundations within a single overriding concept, that of authenticity, defining a test of specific purposes as:

> One in which test content and methods are derived from an analysis of a specific purpose target language use situation, so that test tasks and content are authentically representative of tasks in the target situation, allowing for an interaction between the test taker's language ability and specific purpose content knowledge, on one hand, and the test tasks on the other. Such a test allows us to make inferences about a test taker's capacity to use language in the specific purpose domain (2000:19).



© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-01330-7 - Issues in Testing Business English: The Revision of the Cambridge Business English Certificates
Barry O'Sullivan
Excerpt
More information



This definition highlights the core element of Douglas's view of LSP tests; that of *authenticity*. Douglas does not see this as being a simple matter of replicating specific purpose tasks in a testing context, but of addressing authenticity from two perspectives. The first perspective is that of *situational* authenticity, where LSP test tasks are seen as being 'authentic' in that they are derived from an analysis of the language use domain with which they are associated. The second perspective is *interactional* authenticity, which relates to the actual processing that takes place in task performance, what Weir (2004) refers to as theory-based validity.

This definition has not remained unquestioned. In fact, Douglas (2001) himself acknowledges that there are a number of issues left unanswered by his definition, an argument also made by Elder (2001). This criticism focuses on what Elder (2001) sees as the three principal problematic areas identified in the work of Douglas, namely, the distinguishability of distinct 'specific purpose' contexts; authenticity; and the impact (and interaction) of non-language factors.

By non-language factors one of two things is meant. The first relates to the elements of communication not associated with language – in everyday communication, transferral of message is achieved through a combination of language, cues, signals and symbols. There is a broad literature in psychology on this phenomenon (see for example Brown, Palmeta and Moore 2003, Vargo 1994). The second way of looking at this is the impact of background knowledge, in this case of the business domain, on an individual's ability to perform a particular task, in this case related to an aspect of business communication.

The first of these two perspectives is common across all tests of language production, not solely Language for Specific Purposes (LSP) tests. It is not just related to tests of speaking, where variables such as physical appearance, dress, gestures and posture have all been shown to have an effect on interlocutor perceptions of performance (see for example the work in the area of job interviews of Bordeaux 2002, Chia et al 1998, and Straus, Miles and Levesque 2001), but is also to be seen in tests of writing where handwriting and general presentation skills impact on how writing is evaluated by examiners (see for example Sprouse and Webb 1994, Sweedler-Brown 1992). This aspect of performance assessment is certainly a potential threat to test validity, and is typically dealt with in the development of assessment scales or, more likely, through rater/examiner training.

The latter perspective, the extent to which candidates' background knowledge impacts on his/her test performance is again not associated solely with LSP tests. A test of language for specific purposes is situated, by its very nature, in a specific context, and, also by its very nature, expects (if not demands) of its candidates a knowledge of that context. The literature has shown that background knowledge has a significant and apparently systematic effect on LSP test performance (see for example Alderson and Urquhart 1984, 1985, 1988, Clapham 1996, Steffensen and Joag-Dev 1984). It also appears that as a



© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-01330-7 - Issues in Testing Business English: The Revision of the Cambridge Business English Certificates
Barry O'Sullivan
Excerpt
More information



test becomes more highly specific this effect becomes more acute and it would seem that it is at this extreme that the difficulty in teasing apart language performance and task completion occurs – in other words, in a highly specific test, success on a task is dependent on a successful interplay of language and non-language elements. This feature of highly specific tests at one time led to innovations such as in the General Medical Council's Professional and Linguistic Assessments Board (PLAB) oral test where medics assessed the medical content of ESP tasks and the language examiner commented on the language performance (both informal with patients and formal with professional colleagues, on a generic ELT scale) though specialist lexis etc. remained the domain of the subject specialists.

It can be argued that a test of language for a specific purpose should not even try to avoid the background knowledge issue, as it is this that defines the test. How we deal with the situation will depend on the degree of specificity of the test and the inferences we intend to draw from performance on the test.

Turning to the remaining criticisms of an ESP approach to testing, we can see that there are basically two questions that should be addressed. These are:

1. Distinguishing LSP from general language – is it possible and/or feasible?
2. Authenticity – can LSP tests be made both situationally and interactionally authentic?

## Distinguishing LSP from general English

There is a considerable body of work over the last thirty years which has quite clearly demonstrated the distinguishability of language use in specific contexts. We can point to the work on the definition of language needs and usage in specific contexts of needs analysis researchers and theorists. Among the influential early work were studies undertaken by Hawkey (1978), who offered a practical demonstration of how needs analysis can lead to a specific purpose curriculum, and Alwright and Alwright's (1977) practical advice on an approach to the teaching of medical English.

In the area of testing language for specific purposes, perhaps the most important undertaking was that of the London Chamber of Commerce and Industry Examinations Board (LCCIEB) in 1972. The LCCIEB had been providing business-related qualifications around the world for almost a hundred years when, in 1972, its language section undertook a major analysis of 'foreign' language use involving over 11,500 employees of almost six hundred international firms. This analysis, and the replications undertaken in the Federal Republic of Germany, France, Greece and Spain between 1982 and 1985, were to prove influential in the development of teaching and testing practice in the UK during the 1970s and 1980s.

In a series of seminal articles in the 1980s, Alderson and Urquhart (1984,



© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-01330-7 - Issues in Testing Business English: The Revision of the Cambridge Business English Certificates
Barry O'Sullivan
Excerpt
More information



1985, 1988) found that 'academic background can play an important' though not consistent 'role in test performance' (Alderson and Urquhart 1985:201) and that 'particular groups of students may be disadvantaged by being tested on areas outside their academic field' (Alderson and Urquhart 1988:182). They also suggested that their studies 'demonstrated the need to take account of other factors, such as linguistic proficiency' (Alderson and Urquhart 1985:201). At about the same time Steffensen and Joag-Dev (1984) demonstrated the significant impact on comprehension of a reader's cultural background. The picture that is developing here is that background knowledge is a significant factor in specific purpose language testing, a point that was made by Clapham (1996) with reference to highly specific tests.

In fact, Clapham's (1996) study provided quite a few answers, or at least directions in which to look for answers, to many of the questions asked about the impact of background knowledge on performance in LSP tests. While looking at performance on a test of English for academic purposes (International English Language Testing System IELTS), Clapham's interpretation of the results of her in-depth and complex study have direct consequences for the testing of language for any specific purpose. It is therefore worth looking back over Clapham's work. Among other things, Clapham reports that:

- . . . students achieved significantly higher scores on the module in their own subject area than on the module outside it (1996:188) … [though] the results depend on the specificity of the tests (1996:189)
- . . . it is possible to identify some of the characteristics which lead to passages being more or less specific, but that these characteristics are not always immediately obvious (1996:191) . . . [though] it was the rhetorical function of the passages rather than the sources of the texts which affected their specificity (1996:191)
- it is not always easy to classify candidates into simply defined subgroups, as the evidence from Clapham indicates that her participants were widely read outside of their own area of study (1996:192–3)
- it seems likely that as the modules became more subject specific, background knowledge had a proportionally stronger effect on test scores (1996:193). In addition, subject area familiarity made a significant contribution to test scores, whereas topic familiarity did not . . . [this] suggests that knowledge of a subject area might have a greater effect than topic familiarity on the subject specificity of a reading passage (1996:193)
- there seemed to be a threshold below which students did not make use of this [background] knowledge, and above which they did (1996:194).

The implications of the work referred to earlier in the chapter (e.g. Barber 1962, Hüllen 1981a, 1981b, Johns 1980, Lackstrom, Selinker and Trimble 1973, LCCIEB 1972, Schröder 1981, Selinker and Douglas 1985, Swales 1971, 1984, Weir 1983) when seen in light of these findings suggest that there is a



© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-01330-7 - Issues in Testing Business English: The Revision of the Cambridge Business English Certificates
Barry O'Sullivan
Excerpt
More information



clearly definable language of business (and of other areas of specific interest such as science, technology etc.) and that where tests are devised with a deliberately high level of specificity towards an explicit area, then candidates whose background is grounded in that area can be expected to outperform candidates from a different background, given similar linguistic competence.

There is still a problem, however, in defining the boundaries of specific context areas (Cumming 2001, Davies 2001, Elder 2001). It appears to be the case that while we can identify particular aspects of language use as being specific to a given context (such as vocabulary, syntax, rhetorical organisation), we cannot readily identify exact limits to the language that is used in that context. This is because there are no 'exact limits'. Business language, like scientific or medical language is situated within and interacts with the *general language domain*, a domain that cannot, by its very nature, be rigidly defined.

## Authenticity

Though Douglas (2000) built his definition of what makes a test 'specific' around the notions of situational and interactional authenticity, he later (Douglas 2001) pointed to some difficulties in operationalising such a definition. The notion of situational authenticity is relatively easy to conceptualise. Situational authenticity refers to the accurate reflection in the test design of the conditions of linguistic performance from the language use domain – Weir's (2004) text and task demands. Tests such as that for air traffic controllers described by Teasdale (1994), where candidates were tested in a situation that closely replicated the specific purpose domain, are as close as we can get to a completely situationally authentic test. The mere fact that the event is being used as a test lessens the authenticity – though I'm sure that few readers would expect that the ability of air traffic controllers to cope linguistically with the demands of their work should be tested in a truly authentic situation! The opposite to this would be the relative situational inauthenticity of the MATHSPEAK test, the specific purpose version of the SPEAK (the institutional form of the Test of Spoken English, the TSE) referred to by Elder (2001), where there is no attempt made to replicate the teaching context it is designed to be generalised to.

However, in the case of interactional authenticity there is a lesser degree of certainty in that, to the present time, it has not been clearly conceptualised, let alone operationalised. Though the common view (that the test should result in an interaction between the task and the relevant language ability) is clear enough, to my knowledge there has not been a significant contribution to its operationalisation – that is, insufficient work has been done to link context-based validity elements to theory-based processing. Test developers and researchers tend to rely on anecdotal evidence or 'expert' judgements to make decisions on the interactional authenticity of a test task – in the review of a range of business language tests that comes later in this chapter, I fall foul of the same tendency.



© Cambridge University Press www.cambridge.org

So, critics of an LSP approach to language testing have raised genuine concerns regarding the distinguishability of distinct 'specific purpose' contexts, authenticity, and the impact on test performance of non-language factors – not just for LSP testing but for language testing in general. I do not believe that these are insurmountable and I will return to the matter in the final chapter of this book.

## Assessing performance

While the above issues have focused on the test content and on the theoretical justification for utilising a particular test task, there are other issues in LSP testing that have not really been addressed. Like any test, the reliability (stability, consistency and accuracy) of LSP tests is central to the test's value. In the section devoted to reliability in the context of the BEC suite (Chapter 2) I look in some detail at this issue, so I will not spend time or space here in an extended discussion, except to say that the way we estimate and report the reliability of tests such as the BEC suite is in need of re-appraisal as the statistical approaches taken to date offer us only a limited understanding of the true reliability of these tests.

A related issue is the way in which we evaluate or assess writing and speaking test performances, in that it is associated with the creation of the test score, which is central to any test.

There are a number of issues here:

- the scale criteria
- the level represented by the scale
- the use of the scale (who, how etc.).

### The scale criteria

Though the literature abounds with scales that do not seem to have been derived from any particular theoretical or empirical base, the movement in the 1990s towards more supportable scale development means that the current rating scales which reflect best practice in the area tend to have a sound basis (see North 1996, North and Schneider 1998). While the whole area of rating scale development is far too complex to be dealt with adequately in this short section, it is important to point to the need for any rating scale to be based on the same model or perception of language as drives the rest of the test development process. A good example of this are the rating scales used in the Cambridge ESOL Main Suite examinations (Hawkey 2001).

In their response to the criticisms voiced by Foot (1999), Saville and Hargreaves (1999) present a model of communicative ability upon which the Cambridge ESOL Main Suite speaking examinations are based (see Figure 1.1).



© Cambridge University Press www.cambridge.org

CAMBRIDGE

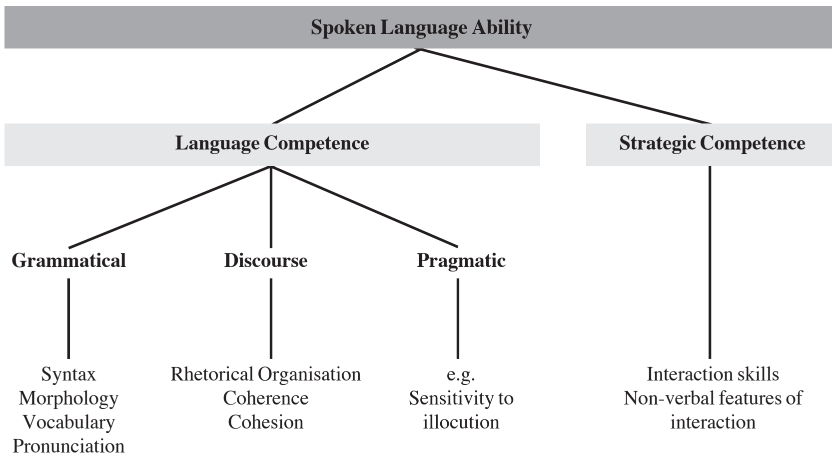
Cambridge University Press
978-0-521-01330-7 - Issues in Testing Business English: The Revision of the Cambridge Business English Certificates
Barry O'Sullivan
Excerpt
More information



This model is based on the earlier work of Canale and Swain (1980) and Bachman (1990), as well as on the Council of Europe specifications for the Waystage and Threshold levels of competence (Saville and Hargreaves 1999:46).

We can see that language competence is described in terms of Bachman (1990:84–98) and Bachman and Palmer's (1996:67) organisational (grammar and discourse), pragmatic and strategic competences.

**Figure 1.1 Communicative language ability**




*Source:* Saville and Hargreaves (1999:45)

The rating scales used in the Cambridge Main Suite Speaking paper examinations consist of four criteria, grammar and vocabulary, discourse management, pronunciation and interactive communication, each of which is awarded a score in the range of 0–5. Though it is not clear from Saville and Hargreaves exactly how the scale is meant to reflect the model of competence they quote, it would appear that it is meant to operate as represented in Figure 1.2.

It is clear from this figure that the notion of pragmatic competence is not explicitly dealt with in the scales (for convenience, only the middle score of 3 is presented in this figure, though the descriptions offered here are similar to the other levels in terms of relevance to model criteria). The notion of pragmatic competence (or knowledge) is seen by Bachman and Palmer as being related to the ability to 'create or interpret discourse by relating utterances or sentences and texts to their meanings' (1996: 69). In other words, pragmatic competence is seen as being comprised of functional and sociolinguistic knowledge and as such has been identified here with the criterion discourse management – which, though the name implies an ability to 'manage' the interaction (in the sense of Bygate 1987), in the context of this scale it is actually concerned with coherence,



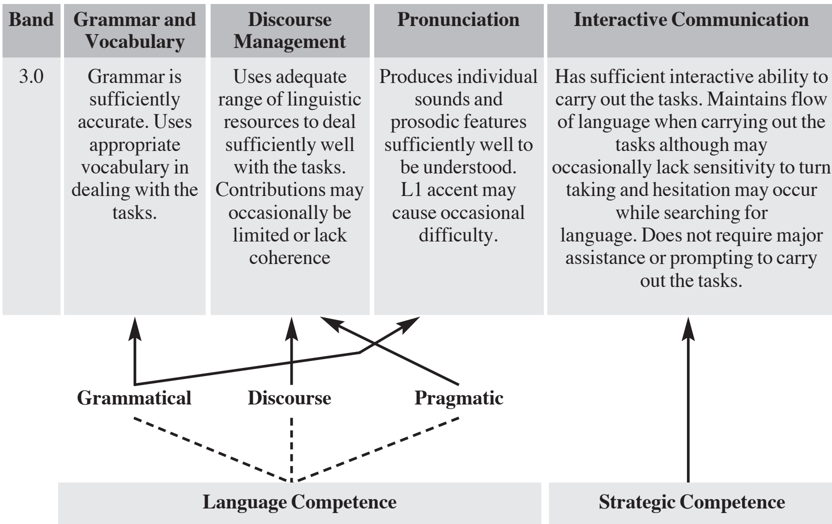
© Cambridge University Press www.cambridge.org

CAMBRIDGE
Cambridge University Press
978-0-521-01330-7 - Issues in Testing Business English: The Revision of the Cambridge Business English Certificates
Barry O'Sullivan
Excerpt
More information



cohesion and, if this representation is accepted, an ability to demonstrate functional and sociolinguistic competence.

**Figure 1.2 Communicative language ability and the Cambridge ESOL FCE analytic scale**



| Band | Grammar and Vocabulary | Discourse Management | Pronunciation | Interactive Communication |
|---|---|---|---|---|
| 3.0 | Grammar is sufficiently accurate. Uses appropriate vocabulary in dealing with the tasks. | Uses adequate range of linguistic resources to deal sufficiently well with the tasks. Contributions may occasionally be limited or lack coherence | Produces individual sounds and prosodic features sufficiently well to be understood. L1 accent may cause occasional difficulty. | Has sufficient interactive ability to carry out the tasks. Maintains flow of language when carrying out the tasks although may occasionally lack sensitivity to turn taking and hesitation may occur while searching for language. Does not require major assistance or prompting to carry out the tasks. |



When advocating a move towards an integrated language/specific area ability approach, Douglas (2000) suggests using what he refers to as 'indigenous' scales in LSP tests. The argument being that the criteria actually employed in the evaluation of specific purpose performances are specific to the context of that performance – a position which is seen as support for the *insepa-rability* of language and performance of specific purpose tasks (Douglas 2001, Elder 2001). While the case made by Douglas is strong, there are a number of points which still need further consideration.

The central problem here is one of construct definition, and therefore of the inferences that are to be drawn from a particular test. In the case of the Occupational English Test (OET), for instance, which is criticised by Douglas and by its principle creator, McNamara (in Jacoby and McNamara 1999) for using a 'general purpose' rating scale, rather than one devised from an analysis of the target language use (TLU) situation, the criticism has some basis, in that the scale used was a rather primitive adaptation of the FSI oral proficiency scale (Wilds 1975). However, the test, for whatever reason (the one suggested was bureaucratic expedience) was meant to offer a measure of the ability of overseas health professionals to cope with the English language demands of their particular medical specialisation. The inferences to be drawn from performance on the test were therefore related to their language competence, nothing else. In



© Cambridge University Press www.cambridge.org