Tasks and criteria in a test of oral communication skills for first-year health science students: where from?

Elisabeth Grove and Annie Brown
The University of Melbourne

Abstract

This paper reports on the development and validation of a specific-purpose language test for medical undergraduates at Melbourne University. It focuses on the advantages of a collaborative approach whereby test developers work closely with representatives of the target context (in this case, medical faculty staff) at all stages of the project, including the design and development of tasks and criteria and the setting of standards. Only by doing this, it is argued, can language specialists ensure that the test reflects not their perspectives on what is relevant, but those of the stakeholders in the testing process.

1. Introduction

Much attention is paid in communicative language testing to designing test tasks which appear to be 'authentic' representations of the 'real world' domain of language use. More often than not, however, there is too little relation between the *tasks*, which lend an appearance of validity to the test, and the assessment *criteria*, often developed *post facto*, against which candidate performance is judged. In addition, where tests of communicative ability for particular contexts or domains of activity are concerned, it is by now well known that the perceptions of communicative quality held by ESL/communications skills specialists do not necessarily reflect, or sit well with, those of professionals within the target context. Language specialists rely heavily their own, language- teaching-derived, criteria whereas professionals from other contexts often privilege aspects of communication which push the boundaries of, or even go beyond, the linguistic. This has been recognised by the descriptive term, 'indigenous' assessment used by Jacoby and McNamara (1999) to identify the assessment orientation of these professionals or 'insiders'. The mismatch between the two is perhaps one of the reasons that language tests are notoriously unreliable indicators of general

Melbourne Papers in Language Testing 2001 Volume 10.1 pp. 37-47. The Language Testing Research Centre, The University of Melbourne.

communicative performance in the target context (Brown 1994, 2000; Davies 1990; Douglas 2000; Jacoby and McNamara 1999; McNamara 1996). If tests of language proficiency for specific purposes are to provide information that is to be genuinely useful[1] to test users, the nature of these differences of perspective must be both recognised and acted upon in the design of test tasks and criteria.

In the project discussed in this paper, the development of a diagnostic test of oral communication skills for first-year undergraduate health sciences[2] students (both native and non-native speakers of English) at the University of Melbourne, we therefore adopted a collaborative approach: the 'clients', staff of the three health sciences schools, were closely involved in all aspects of the project, from needs analysis to validation. The paper reports on the collaboration between the test developers and medical educators to determine the skills which the test needed to tap, and thus to establish the criteria which informed the design of the tasks. The negotiations between the two groups of participants involved accommodating differing views of communication. For us as test developers, this meant crossing the boundaries set for more general tests of second language proficiency, and for the medical specialists, it meant making explicit and calling into question what was actually understood by the term 'communication' within the medical field. These negotiations resulted in a test that was felt to be innovative, flexible and useful.

2. Background to the test

The importance of effective communication between patient and health professional is now widely recognised as being of vital importance. Health professionals need not only ever more complex technical knowledge but also high levels of communicative ability. In Australia, as elsewhere, medicine, dentistry and physiotherapy are high-status occupations, requiring top scores for entry to prestigious courses of study. They attract students who typically excel in the sciences, but too often, high achievement in these areas of study is not matched by equivalent communication and interpersonal skills. In recent years, the most significant challenge to the education of health professionals in Australia has been the increasing number of

---

[1] For a definition of what test 'usefulness' might be expected to encompass, see Bachman and Palmer 1996.

[2] Health sciences in this context includes medicine, dental science and physiotherapy.

'international' students, who currently comprise more than one third of the total enrolment in the Medical School at Melbourne University. For this group of students in particular, there is concern among medical educators that many lack the communication skills necessary for success in their studies and future careers.

Recent changes to the health sciences curricula at the University have also heightened the importance of effective communication. Until 1999, Melbourne medical students followed a traditional six-year course of study, in which clinical placements involving direct contact with patients did not take place until the fourth year, after three years of entirely university-based academic study. In the new curriculum, clinical studies are introduced at the beginning of the course. The new curriculum also centres on an integrated teaching/learning model, known as Problem-Based Learning (PBL), which emphasises group work, shared problem-solving and independent learning. Health Sciences staff expected that these two changes together would make greater demands on the communication skills of students from the beginning of their university studies. The Faculty therefore decided that a test providing diagnostic information was needed from the outset. The purpose of the test was:

a) to identify students with particular (especially ESL or cultural) problems and refer them to concurrent support as early as possible, and

b) to identify those whose communication skills might put them at risk of performing below their potential, so that course tutors could 'keep an eye on them'.

3. Needs analysis

The first task was to identify exactly what sort of information the medical educators needed and what kinds of communication problems they found in their students. By means of the now-familiar techniques of LSP needs analysis (eg. Alderson *et al.* 1995, Davies 1990, McNamara 1996), we hoped to extract the relevant aspects of performance thought to underlie academic and clinical communication tasks, particularly those which students are known to find difficult, and thus to determine both test content and assessment criteria. Data were collected by a variety of means: interviews with a broad range of clinical and academic educators; direct observation of students practising clinical assessment tasks and case presentations; scrutiny of teaching materials; and a survey questionnaire of students who had sought concurrent support in language and communications during the year.

The needs analysis generated a huge list of concerns, what Alderson *et al* (1995: 23), have politely called "a large taxonomy of variables", about the students' communication skills, as well as identifying the tasks and language functions required of them, from first to final year and beyond. The tasks considered particularly important or difficult included, in the case of professional/clinical communication, eliciting a case history, making a case presentation, and justifying a diagnosis and treatment plan. In the case of the curriculum-related aspects of the course, there were concerns about participation in group work, seminar presentation skills, and so on. The skills in which many students were found lacking ranged from understanding idiomatic Australian English to possessing an appropriate bedside manner. In particular, there was recurrent concern about 'cross-cultural' issues (such as inappropriate tone or register in talking to patients, or unsuitable attitudes, for instance, towards explaining procedures). It soon became evident to us that the medical specialists considered the behavioural and attitudinal aspects of clinical and academic interaction to be at least, if not more, important than any linguistic difficulties that students may have. In order to incorporate their concerns, and to make sure we really could identify students likely to be at risk for reasons both linguistic and non-linguistic, we realised that it would be necessary to move closer to the medical model of communication than was initially felt to be comfortable.

However, finding out what that model consisted of was by no means straightforward. In our search for a theoretical model of communication broader than that of second language proficiency testing, we scrutinised the literature and teaching materials on medical communication. While there was much emphasis on the importance of communication, they offered no explicitly theorised model, but rather, implied one. The perspective of clinical educators emerged as essentially *task-based* and *instrumental*; that is, conceptualised in terms of the *problems* involved in communicating with patients and peers within these tasks, and specific situation-specific behavioural *strategies* (such as what to do if the patient is loquacious or silent) for dealing with such problems as effectively as possible. This reflected the outcomes of the needs analysis conducted at the university.

What the medical educators wanted was a test that could diagnose students' needs in relation to the types of tasks they felt presented special difficulty: the behaviours or problems included both linguistic and non-linguistic aspects of performance. As language testers, we faced a dilemma. If the test were to provide useful diagnostic information, it would need to address some of these non-linguistic issues, and this is not a comfortable model for second language

testers used to working with theoretical linguistic models of traits and abilities. Recognising the diagnostic rather than gate-keeping purpose of the test, we agreed to include not only the linguistic skills we took to underpin successful performance of key communication tasks, but also to comment, albeit tentatively, on some of the behavioural aspects of the candidates' performance - which entailed interpreting problematic aspects of task performance in terms of cultural background, personality, maturity, familiarity with test task, nervousness, and so on, in order that appropriate action could be taken.

4. Test tasks

The vast 'worry-list' of problems uncovered by the needs analysis had to be radically culled in order to arrive at a manageable test design which would take both task and assessment requirements into account. It was decided to include tasks which elicited behaviours relevant to both the clinical and the academic context. Because the students were only at the beginning of their courses, unlike many other tests (particularly tests of *professional* language competence), it would not be possible to simulate the target context and require students to perform simulated professional tasks or display specialised subject knowledge. For this reason, the content of two tasks was devised to relate to the course of study but without trying to mimic it. Two tasks were decided upon - a monologic presentation and a dialogic discussion. The monologue allows an assessment to be made of presentation-related skills relevant both to the clinical context (the case presentation) and to seminar work (class presentation), both of which require a high level of skill in synthesising and organising complex material, as well as clarity and precision of expression. The dialogue focuses on the interactional and interpersonal skills also relevant to the two contexts (that is, discussion and negotiation in seminar and in professional-client interaction).

In Task 1 (see Figure 1), students listen to a 10-minute audio-taped tutorial discussion of a non-specialised but medically-oriented topic ('the ethics of medical experimentation on human subjects') and then present a 4-minute oral summary. Unlike more conventional EAP tests, where the stimulus typically involves an extract from a lecture (by its very nature already structured), in this case, the stimulus consists of dispersed information and conflicting points of view which the candidate is required to synthesise and structure for a listener. It is important to note that the tasks and the test itself were not designed *solely* for non-native speakers, but rather, are the types of skills expected of *all* students. Native speakers and non-native speakers alike find this first task particularly strenuous, but it is

considerably more so for many non-native speakers, who have a much higher rate of unsatisfactory performance.

The second task, which involves informal discussion of a general topic relating to education, aims to elicit the types of skills typically required in undergraduate studies – explaining, presenting and defending opinions, hypothesising and speculating. It also relates to the communication skills required in the clinical encounter – namely, explaining, justifying, persuading, and so on. In addition, interpersonal skills are of particular salience in clinical communication with patients. Again, these skills (both functional and interpersonal) are not necessarily exhibited by all native speakers.

Figure 1: Test tasks

Task 1	Task 2
Summary presentation (monologue)	**Discussion (dialogue)**
Clinical context Case presentation	*Clinical context* Consultation
Academic context Seminar presentation	*Academic context* Problem-based learning (PBL) Group discussion and problem-solving
"The ethics of medical experimentation on human subjects"	"Education"

5. Assessment criteria

As noted earlier, the criteria for tests of language ability are often developed *post hoc* and by test designers, who are by and large language specialists. As a consequence, the criteria they arrive at may not necessarily reflect the criteria relevant to subject specialists or in the real-world context. In developing this test, we were very much concerned that the criteria provide information on students' communication which reflected the sorts of information the medical faculty staff had indicated as important to them. In other words, we needed both linguistic and task-related criteria, as well as information on cultural and personal communication traits.

To this end, the tasks were designed with the sorts of judgements in mind that had been identified in the needs analysis. A draft set of criteria by which to judge performance was developed at the same time as the tasks. Feedback on these criteria was then provided by the committee of medical specialists whose role it was to oversee the test

development (and associated student support program). The criteria were of two types, what we have termed 'language-related' criteria and 'task-related' criteria. The linguistic criteria (see Figure 2) were concerned with features of ability common to most tests of second language speaking, and one global assessment was made for both tasks[3].

Figure 2: Assessment criteria

Linguistic criteria (both tasks)	
Language	Range of structure & vocabulary Breadth and precision of expression Accuracy
Production	Pronunciation Intonation, stress and rhythm Voice quality
Task-related criteria: Task 1	
Organisation	Macro structure of presentation
Content	Sufficiency and appropriateness
Style	Level of formality, tone Nonverbal behaviour
Fluency & Coherence	Sequencing, linking, clarity of ideas Fluency of presentation
Comprehension of input	
Task-related criteria: Task 2	
Adequacy of participation	Maintenance of interaction Initiative, expansiveness
Quality of ideas	Maturity or quality of thought
Interpersonal skills	Engagement, rapport Nonverbal behaviour
Coherence & expression	Clarity of ideas Cohesion and coherence
Register & tone	Level of formality Politeness Directness Tone of voice

[3] In trialing the test, it was found that the quality of these linguistic features was fairly constant across both tasks.

The task-related criteria, on the other hand, attempt to distinguish between the skills essential to effective communication in each type of task. They were drafted initially on the basis of what the medical staff had identified as key aspects of performance, and later refined on the basis of an examination of student performances, as we attempted to define what it was about particular performances that was satisfactory or unsatisfactory.

In Task 1, the presentation, these include such features as ability to structure the summary presentation effectively, the selection and appropriateness of content, style (including non-verbal features of performance such as eye contact or lack of it, excessive casualness, or individual quirks of behaviour), and so on. In Task 2, the discussion, the emphasis is on skills of participation, quality of ideas, including maturity, interpersonal skills, appropriateness of register and tone, etc. This array of criteria is somewhat broader than those generally found in tests of second language oral proficiency, even tests which are claimed to predict the ability of candidates to perform communication tasks related to academic study or professional contexts. In most such tests, in our experience, other than the narrowly 'linguistic' criteria, assessors are typically required to judge candidates according to a relatively undefined category of 'communicative effectiveness', as though what constitutes 'effectiveness' were self-evident (Brown, 2000). In this test, in contrast, an attempt was made to define the components of task-performance and to operationalise them as explicit and discrete assessment criteria.

Whilst Figure 2 shows the criteria as they appear on the score sheet, raters are, in fact, provided with a much more expanded description of the types of behaviours (and examples) relevant to each category which they can refer to. This explanatory data was compiled, following the test trials, through a process of gathering both language specialists and subject specialists together to review the taped trial performances. In addition, because of the diagnostic purpose of the test[4], for each assessment category, raters are encouraged not only to comment on specific behaviours, but also to comment on why they think these behaviours arise.

Although a host of factors—cultural, interpersonal, cognitive, experiential and institutional, to name just a few—are known to

[4] All students below a certain level are referred automatically to the faculty-based ESL specialist, who is provided with all score-sheets and video-tapes.

influence communicative effectiveness, there has been little effort to acknowledge this in the rating procedure, to give raters a means of dealing with what they know to be non-language-related performance factors. Yet from the few studies of verbal protocols of performance on 'communicative' speaking tests (such as Brown 2000), we know that raters do grapple with the issue of whether particular perceptions (of nervousness or lack of test familiarity, for example) are 'relevant' to their judgements. By asking raters to comment on their perceptions of the underlying reasons for performance, we hoped not only to provide useful diagnostic information about students to the Medical Faculty, but also make the raters feel more comfortable about the 'numbers' they were assigning to students. That is, we provided a way for them to record different reasons for the same scores, something that is not possible in tests which are neither explicit about the specific behaviours of interest ('communicates effectively') nor recognise that factors other than language can affect performance.

But what of the value of this information to the medical educators? What evidence do we have of validity? We made the following attempts to evaluate the usefulness of the test, during the test development process itself, and after its implementation in terms of the information supplied in its first year of operation.

First, before implementation, a benchmarking exercise was carried out where a selection of tapes from the trials were reviewed by a group of medical staff who were particularly concerned with communication-skills issues. They knew many of the first year students who had been involved in the trials (and a particular effort had been made to include students identified by staff as being 'at risk'), so it was possible to compare the assessments and judgements made on the basis of the test criteria with their perspectives on the students. We found that the assessments generally matched, and that the scale worked well in distinguishing those students who needed additional support in both their second language skills and in 'culturally' appropriate behaviour.

On the basis of the benchmarking exercise, the following reporting levels were established:

- Scores of 5 and 6 on the linguistic and task criteria were considered to be satisfactory / not at risk
- Scores of 3 and 2 indicated that students had major problems and were likely to need extra-curricular support.

- Scores of 4 indicated minor problems which were felt not to require instant remediation (especially given the limited ESL support available within the faculty), but which warranted students' on-course performance, particularly in Semester 1, being monitored.

Secondly, after implementation, we followed up with ESL and academic staff whether those students who were considered to be 'at risk' or in need of extra support during their first semester of study had actually been identified as such by the test. This was found to be the case.

6. Conclusion

In reporting on the development of the Health Sciences Communication Skills test, we have attempted to show how, in order to develop a test that would truly be useful, we had to expand our notions of what could reasonably be included to encompass not only second-language specific criteria, but also task-related aspects of performance which may be problematic for both second language and first language speakers. So the criteria included both general linguistic and task-specific criteria. In addition, in order to ensure that appropriate action could be taken (which included not simply ESL support, but attention to cultural and behavioural problems which might impact on students' performance in the course of study) we had to take account of reasons for poor performance on the test. For this reason, raters were encouraged to comment on the specifics of any one candidate's performance *and* to extrapolate about the causes of poor performance. This involved the test developers in being more explicit than is normally the case in second language tests, however, 'communicative', about reasons underlying poor task performance.

References

Alderson, J.C., C. Clapham & D. Wall. 1995. *Language Test Construction and Evaluation.* Cambridge: Cambridge University Press.

Bachman, L. & A. Palmer. 1996. *Language Testing in Practice.* Oxford: Oxford University Press.

Brown, A. 2000. An investigation of the rating process in the IELTS Speaking Module. In R. Tulloh (ed.), *Research Reports Vol. 3.* Sydney: ELICOS: 49–85.

Brown, A. 1994. LSP testing: The role of linguistic and real-world criteria. In R. Khoo (ed.), *LSP: Problems and Prospects*. Singapore: SEAMEO Regional Language Centre: 202–215.

Davies, A. 1990. *Principles of Language Testing*. Oxford: Basil Blackwell.

Douglas, D. 2000. *Assessing Language for Specific Purposes*. Cambridge: Cambridge University Press.

Jacoby, S. & T. McNamara 1999. Locating competence. *English For Specific Purposes*, 18, 3: 213–241.

McNamara, T. 1996. *Measuring Second Language Performance*. London: Longman.