Problematizing content validity: the Occupational English Test (OET) as a measure of medical communication[1]

Tim McNamara
Department of Linguistics and Applied Linguistics
The University of Melbourne

Abstract

In the Occupational English Test, an ESP test for health professionals, as in other workplace performance tests, selected workplace tasks are simulated. However the difficulty of simulating authentic communicative process under test conditions, and the inseparability within authentic performance of professional knowledge and language behaviour, suggest that content validity may be only superficially achieved.

1. Introduction

In this paper, the stages of the development of an occupationally related performance test of the *work sample* type will be outlined, through an extended example, the development of the Occupational English Test (OET), a test of ESL for health professionals. The term *work sample* is used here in a broad sense and includes tests which involve simulation of any kind, particularly where the content of the simulation is based on an analysis of tasks and roles in the criterion (Figure 1).

The establishment of appropriate test content is of particular concern in work sample tests, and the account of the test development process will illustrate how it was attempted in this case. The replicable aspects of the criterion can in principle include: task stimulus; task demand; task processing; and the criteria by which successful performance of the task is evaluated. However, the discussion will focus on the limits to the ways in which simulation of the criterion is in fact possible in work sample tests,

[1] Material in this chapter draws heavily on McNamara (1996) ch 4. Thanks are given to the publishers, Addison-Wesley Longman, for granting permission to reproduce this material.

Melbourne Papers in Language Testing 1997 Volume 6.1 pp. 19-43. The Language Testing Research Centre, The University of Melbourne.

thereby raising a number of complex questions about the nature of performance in performance assessment, and about the validity of the work sample approach to language performance testing.

Figure 1. Test and criterion in work sample tests

Test		Criterion
a performance or series of performances, simulating / representing the criterion	⇐ modelling and sampling ⇐ ⇒ inferences about ⇒	A series of performances subsequent to the test; the target
(observed)		(unobservable)

2. Stages in determining the content of a work sample performance test

In work sample tests, an attempt is made to sample test content from the communicative tasks facing the candidate in the target language use situation. The design of the test will therefore involve research on the communicative demands of the target setting. This research is likely to have four components, as follows. (A minimum of detail will be provided in the description of each of these components, as the extended example on the OET in the following section will illustrate more clearly).

Literature search

It is necessary to establish what research has already been carried out on the communicative demands of the target setting. For example, relevant to EAP test design, a body of research exists on the communicative tasks facing students at universities in the UK (Weir, 1983a; Weir, 1983b; Emmett, 1985) and in North America (Powers, 1985). In job-specific assessment, less may be known.

Consultation with expert informants

People responsible for professional education and training for the workplace are likely to have a more explicitly articulated view of the nature of the workplace and its demands than those engaged in the work less reflectively. Such expertise constitutes a resource for the test designer in the process of task specification. Others such as work supervisors may also serve as informants. Interviews with these informants may help at the stage of an initial orientation to the workplace, particularly if it is relatively unfamiliar to the researcher. Informants may help in establishing a preliminary inventory of tasks, to be examined critically at the next stage of the process (observation and job analysis). They may also be a useful sounding board as the researcher more clearly formulates his/her understanding in the course of the investigation. The views of such informants should not of course simply be accepted at face value, but considered in the light of evidence from other sources. Expert informants may also be involved in the subsequent preparation and validation of test materials (see below).

Job analysis and workplace observation

This stage involves categorisation of communicative tasks in the workplace (or other target language use situation) based on observation, interviews, questionnaires/surveys, and other methods. The complexity of this stage has often been underestimated, and the range of techniques used rather limited (Mawer, 1992). Morgan (1994) for example argues convincingly that the assessment of the role of literacy tasks in the workplace in accordance with recent Australian government policy has been naive and misguided, as it has been conducted by ESL teachers with little understanding of the culture of industry and with inadequate training in methods of data collection in such settings. She suggests a range of techniques that might be used.

Collection and examination of texts (spoken and written) from the workplace.

As part of the empirical investigation in (c) above, texts may be collected and examined for complexity, length, discourse organisation and other linguistic features. Material from these texts may appear in the test itself, although the inauthenticity of the conditions of production and reception of texts in the testing situation compromises the authenticity of such texts as

representative of workplace communication tasks. This will be discussed further below.

3. An example: the case of the Occupational English Test (OET)

In this section, we will consider how the above process for the development and validation of test content were carried out in the case of the Occupational English Test (OET), an Australian Government test of ESL for health professionals.

3.1 Test background

As part of its annual intake of immigrants and refugees, some hundreds of overseas-trained health professionals are currently entering Australia each year as permanent residents. The majority of these are medical practitioners, but a number of other health professional groups (nurses, dentists, physiotherapists, occupational therapists, speech pathologists, veterinary surgeons and several others) are also represented. The process of registration for practice in Australia typically involves the following three steps after an initial verification of documentation:

(1) The Occupational English Test, an English language proficiency test taken by each of the professions involved; its development and administration are in the hands of the National Languages and Literacy Institute of Australia (NLLIA) on behalf of the National Office for Overseas Skills Recognition.

(2) Profession-specific pencil-and-paper tests of professional clinical knowledge, developed by the relevant professional examining body, for example the Australian Medical Council.

(3) Performance-based tests of clinical competence, again conducted by the relevant professional examining bodies.

The first two stages may be completed before entry to Australia, and official policy encourages this, although in fact only a small percentage of those seeking registration do so (for example, approximately ten per cent of candidates for the OET sit outside Australia). Passage through the three stages may take some considerable time - a period of up to two years is quite common.

Prior to 1987, the OET was a test of general English proficiency, and was attracting increasing criticism from test takers and test users in terms of its validity and reliability. In response to this, a series of consultancies was initiated on reform of the test. The report on the first of these, which was carried out by a team at Lancaster University in the United Kingdom, recommended the creation of a test which would 'assess the ability of candidates to communicate effectively in the workplace' (Alderson *et al.* 1986: 3). A series of further consultancies (McNamara 1987, 1988, 1989a) attempted to operationalize this recommendation by establishing the format of the new test and developing and trialing materials for it.

Steps 2 and 3 for registration are demanding for those health professionals whose clinical experience is restricted to contexts rather different from Australia, a country with a technically sophisticated health care system, where diseases are linked to the lifestyle of a relatively affluent industrial and post-industrial society. This is true for a majority of candidates, who most frequently come from countries in Eastern Europe, the Middle East, the Indian sub-continent and South-East Asia. In order to have a chance of success on these practical, clinically based tests, reflecting roughly the standard of final year medical training in Australia, candidates must have some access to experience of clinical practice in Australia. A limited number of places on hospital based bridging programs is available.

Thus, it was recognised that an important function of the reformed Occupational English Test would be to serve as a screening for entry to a setting providing supervised clinical familiarisation. Experience of the clinical setting, no matter how informally or in however limited a way, is indispensable for success in the clinical examinations. Institutions who conduct such programs felt that a screening of candidates' ability to cope linguistically with bridging programs was a necessary function of the test.

3.2 Resources/constraints

As indicated above, the OET was initially under-resourced, but eventually the Australian Government responded to repeated protests about its unsatisfactory status, initiated consultancies on its reform and has since borne the cost of the routine development of appropriate test materials (2-3 new forms per year) and test administration. As a relatively specialised performance test, the

OET is expensive to operate, but its function is important politically and socially; as a result, resources have been available for a performance testing approach involving specialisation of materials for different health professions. Administratively, there is a preference for separate reporting of sub-scores in each of the macroskills. The main constraint on the test is a legal one, namely that the testing of English language *communication* skills must be entirely separate from the testing of clinical aspects of professional competence carried out by assessment panels specific to the health professions concerned. (This contrasts with the equivalent test in the UK, the PLAB test.)

3.3 Content selection

Although content selection has been seen as crucial in work sample tests, there is a debate on the status of content selection in relation to test validity. Messick (1989: 17) takes the position that 'in a fundamental sense so-called content validity does not count as validity at all'. This position has not been reflected in thinking in language testing, where there has been a great emphasis on content validation, particularly in the important British tradition of EAP testing. Davies (1984), for example, sees the stage in which the content and the general layout of the test are planned and the type of test item is decided on as the crucial stage in the development of such tests.

The establishment of content validity, according to Davies (1977: 62), involves the following:

> *An assessment must be made of just what the learners whose proficiency is to be tested need to do with the language, what varieties they must employ and in what situations they must use those varieties. This is an arduous task, and one based largely on guesswork, but it can be intelligent guesswork and it is essential for constructing a proficiency test.*

It is not entirely clear what motivates Davies's scepticism here. A number of procedures were employed to guide the 'guesswork' in the case of the OET: these included *interviews* with those involved in the professional education and training of both overseas-trained and locally-trained health professionals, the administration of a *questionnaire* to those with direct experience of the relevant workplace roles, *direct observation* of the workplace, and *analysis*

of available characterisations (both within and outside applied linguistics) of key communication tasks.

3.3.1 *Consultation with expert informants*

Relevant expertise was to be found among 3 groups of informants:

a) those responsible for professional education in clinical settings in each of the professions concerned; of particular help were those who had had direct experience of clinical supervision of overseas-trained health professionals;

b) overseas-trained graduates with some experience of bridging programs in clinical settings in Australia;

c) ESL teachers contributing to such programs (such teachers observe interactions between the graduates and hospital staff and patients in the clinical setting, and base assistance on what they observe to be the communicative difficulties experienced).

In the first part of this data-gathering stage, the views of those concerned with the education of doctors were sought, as doctors are the largest group taking the test.[2] Interviews were conducted, some observation of the workplace was conducted, and a tentative list of relevant workplace communicative tasks was drawn up; this was then discussed with the informants in group (a) and some of the informants in group (b) above. A questionnaire was then developed which attempted, following Weir (1983), to identify those communication tasks facing graduates in clinical bridging programs which were perceived to be most *frequent*, and those which were seen to be most *complex* or *difficult* (see also Candlin, Leather and Bruton, 1984).

The questionnaire was administered to 42 overseas medical graduates in three areas of Australia who had current or recent experience of the hospital setting in Australia in an observer or trainee role. The graduates were asked to estimate the *frequency* of contact with various other personnel in a variety of channels (face-

---

[2]The resources of time and money available meant that the time spent on the work of doctors could not be repeated for each of the professions concerned. This is one among many examples of the practical difficulty of dealing with disparate work settings in a single test. Of course, even for doctors, work settings will differ enormously, and assumptions about prototypicality had to be made.

to face, telephone, reading/writing). A five-point scale was used to measure frequency, with 0 representing 'no contact' and 5 representing 'very frequent contact'. Table 1 gives details of the twenty commonest communication tasks for doctors, based on an analysis of 42 completed questionnaires.

The results show that the main communication tasks facing overseas-trained doctors in the hospital setting involved mainly oral communication skills. The ten most frequent communication tasks were oral: the most frequent task was face-to-face communication with patients, the next 5 ranks being taken up by face-to-face communication tasks involving hospital personnel.

Table 1 Perceived frequency of communication tasks

Rank	Mean frequ-ency	Channel	Person(s)
1	4.24	Face-to-face	Patient
2	4.07	Face-to-face	Registrar/resident
3	3.76	Face-to-face	Other foreign graduate
4	3.40	Face-to-face	Consultant
5	3.10	Face-to-face	Nursing staff
6	3.00	Face-to-face	Medical student
7	2.45	Telephone	Telephonist
8	2.30	Telephone	Other foreign graduate
9	2.26	Face-to-face	Receptionist/secretary
10	2.15	Telephone	Registrar/resident
11	2.13	Reading/writing	Registrar/resident
12	2.00	Telephone	Receptionist/secretary
13	1.93	Face-to-face	Radiologist
14	1.75	Telephone	Nurse
15	1.68	Reading/writing	Radiologist
16	1.60	Face-to-face	Hospital administration
16	1.60	Face-to-face	Patient's relative
18	1.58	Reading/writing	G.P.
19	1.53	Reading/writing	Nurse
20	1.48	Reading/writing	Consultant

The graduates were also asked to gauge the relative *complexity* or *difficulty* of specific communication tasks for doctors with limited

English skills, using a scale from 0 'not difficult' to 5 'extremely complex or difficult'. Table 2 gives details of the results obtained.

Several tasks relating to aspects of oral communication with *patients* were seen as complex; one aspect of oral communication with other doctors was also singled out (*case presentation*). Interestingly, other aspects of inter-professional oral communication did not seem to present great difficulties (*understanding colleagues' language on ward rounds* and *understanding case discussions*), despite their perceived frequency, perhaps because of the greater degree of shared 'knowledge in the case of communication among health professionals compared with doctor-patient communication.

Table 2 Perceived complexity or difficulty of communication tasks

Rank	Mean complexity	Nature of task
1	3.14	Understanding colloquial language from patients
2	2.63	Understanding local cultural references in order to make judgements about patients' lifestyles
2	2.63	Case presentation
4	2.33	Explaining medical ideas in easy language for patients
5	2.21	Taking a case history
6	2.10	Asking patients questions
6	2.10	Writing letters/reports
8	2.05	Speaking on the telephone
9	2.02	Reading handwritten notes/letters etc.
10	1.88	Clarifying symptoms
11	1.83	Taking part in a group tutorial
11	1.83	Dealing with patients' families
13	1.75	Socialising with other staff
14	1.52	Getting cooperation from patients
15	1.51	Dealing with administration
16	1.40	Dealing with other non-professional hospital staff
17	1.36	Understanding colleagues' language on ward rounds
18	1.33	Getting cooperation from nurses
19	1.30	Dealing with outside personnel
20	1.24	Understanding case discussions

Writing tasks were less frequent, and perceived to be of moderate difficulty.

3.3.2 *Literature search*

The extensive literature on medical communication was examined and evaluated for its relevance to the characterisation of the consultation. An issue emerged from the contrasting perspectives of this extensive research. On the one hand, there were studies which characterised the consultation as an event, with stages, organised temporally and sequentially. On the other, several other broad traditions of research examined the processes in which the participants were engaged, for example from the points of view of social psychology, conversation analysis or detailed discourse analysis. A decision had to be made about what in this literature should inform the content specifications for the test. As it was obviously easier to specify stages of the consultation, some of which could in a rather obvious (and perhaps superficial) sense be simulated in the task design, rather than to attempt to specify the replication of communicative processes, themselves often only revealed by extensive analytical effort, a decision was made to specify test content at the macro rather than the micro level. The limited available resources, particularly in terms of time, were a further constraint. But the implications of this in terms of test validity are problematic. On the other hand, it was not clear whether it was possible to simulate process under test conditions; as we noted above, authenticity of process is a major problem in performance assessment in the work sample tradition (more on this below).

Of the studies carried out by discourse analysts, the most helpful in defining the stages of the consultation were those by Coulthard and Ashby (1976) and Skopek (1979); the work of Candlin, Leather and Bruton (1984) was also particularly useful. Most important of all were studies of the medical consultation conducted from within the medical profession itself, particularly the study by Byrne and Long (1976) of the structure of some 2500 medical interviews conducted by general practitioners.

3.3.3 *Direct observation of the workplace and job analysis*

In the light of the questionnaire results, and the obvious importance of the consultation, time was spent observing workplace

communication in all of the professions served by the test. Throughout, an attempt was made to establish commonalities among the professions as the basis for test task design.

The centrality of *talk* in the work of each of the professions was readily established. Observation confirmed that the consultation (known by several names across the various professions) was common to all. The purpose of the consultation is *assessment* of the patient; physiotherapists and occupational therapists, for example, are expected to make a detailed assessment of the patient before any treatment is given. This assessment is independent of any that might be available from other health professionals involved in the care of the patient, although records of these other assessments also form an important part of the final assessment by each professional.

Consultations between professionals and their patients or clients were observed to contain the following elements[3] which were common to all the professions, although not all of them might be present in every consultation:

1. Assessment of the patient ('subjective assessment') including history taking.
2. Physical examination.
3. Explanation to patient of diagnosis and prognosis and course of treatment.
4. Treatment.
5. Patient/client/relative education and counselling.

Such commonalities provided the basis for a common format for assessment of *speaking* and *listening* skills (see below).[4]

Frequent and important *reading* tasks fell under two headings: keeping up with the professional literature; and reading case notes. As for *writing*, the letter of referral was found to be common to all the professions. When responsibility for care of a patient or client is

[3]Corresponding broadly to the stages identified by Byrne and Long (1976: see above) in consultations involving GPs.

[4]Note that some of these elements (for example #5) are quite complex in terms of their generic structure (Martin, 1984; Swales, 1990; Paltridge, 1995); this brings us back to the problem of working at the macrolevel in characterizing the consultation as a basis for decisions about test content.

handed over to another health professional (usually within the same profession, although not always) a short letter may be written summarising the main facts of the case and the stage treatment has reached, and making a suggestion for further assessment or treatment. Where necessary, the writer bases the information in the letter on records such as case history notes, letters from other professionals concerned in the case, and other documents in the medical record of the patient. Of course, the above characterisation of this writing task is an idealisation; in fact letters may be written without reference to notes, often at the time of the referring consultation, and the letter itself may be a pro forma note. Nevertheless, informants by and large supported the inclusion of a more elaborate and formal version of this writing task in the test, perhaps as much guided by views of appropriate test content and format as by fidelity to the reality of practice.

3.3.4 *Collection and examination of texts from the workplace*

Numerous audiotaped recordings were made, including of case conferences and ward rounds; videotapes of authentic doctor-patient consultations were also available. Examples of written texts were collected or copied. These materials were considered in the light of possible test formats. For example, it had been hoped that a recording of a case conference might be used as the basis for part of the listening test, but the question of confidentiality arose, and it was decided to re-record the case conference from a transcription, using actors. The results proved unconvincing, because of the lack of familiarity of actors in using transcripts of actual spoken interaction (involving overlap, incomplete utterances, etc) as the script from which they were to work.

3.4 Development of specifications

As always with specific purpose performance tests, compromises had to be made about the degree to which it was practicable to tailor materials for particular professions[5]. It was decided to have single sub-tests of Listening and Reading while providing profession-specific content within a common format for the Speaking and Writing sub-tests.

[5]Leaving aside the potential diversity *within* professions.

The broad specifications for the test were thus as follows:

Speaking

Profession-specific content within a common format. 15 minutes. Role play-based interaction[6]: candidate in own professional role, interlocutor in role of patient (client) or relative of patient (client). Two role plays, plus short (unassessed) interview. Assessment in six categories, using rating scale format scoring grid of the semantic differential type.

Writing

Profession-specific content within a common format. 40 minutes. Letter of referral (12-15 lines) based on case notes or extracts from medical records. Assessment in five categories, using rating scale format scoring grid of the semantic differential type.

Listening

Common to all professions. 50 minutes. Short answer question format. Two texts:

(a) talk on a professionally relevant subject;

(b) consultation between a general practitioner and a patient.

Reading

Common to all professions. 40 minutes. Multiple-choice question format[7]. Articles from professional journals.[8]

3.4.1 Development of a scoring system

A procedure used in the original FSI speaking test (Wilds 1975) was used as the basis for the development of a set of scoring categories

---

[6]Role plays (cf Burton's proposals in Candlin, Burton and Coleman 1980) allowed simulation of components 3 and 5 of the consultation, and aspects of 1, listed in section 3.3.3 above.

[7]The MCQ format itself represents a reading task which simulates the reading task facing the health professionals in the tests of clinical competence subsequent to the OET which are in MCQ format. Some informants recommended the incorporation of MCQ format in the OET reading sub-test as it was not clear if people who were doing poorly on the clinical MCQ exam were experiencing problems handling the MCQ format or clinical knowledge problems.

[8]The writing sub-test incorporates a substantial reading task, that of comprehending medical records, usually maintained in note form and containing standard abbreviations. Such records form the stimulus for the writing task where they provide the medical content to be included in the letter of referral.

for the Speaking sub-test (cf Appendix A), and subsequently for the Writing sub-test (cf Appendix B).

For *speaking*, raters assess the performance on each role play on six dimensions separately, and mark their ratings on scales presented in a type of semantic differential format, with anchor terms at either end. In addition, raters are given guidance as to the interpretation of different score points along the scale, described in terms of how they relate to the minimum level of proficiency required of a participant in a clinically based bridging program.

The assessment categories are defined communicatively. For example, where the original FSI scoring system had categories for *Accent*, *Grammar* and *Vocabulary*, these were redefined as *Intelligibility* and *Resources of grammar and expression*. The communicative categories of *Fluency* and *Comprehension* (of interlocutor's input) were retained, and another (*Appropriateness*) introduced. An overall impression category, *Overall communicative effectiveness*, was introduced, and weighted more heavily: scores in this category are added to the average of the scores in the other five categories to achieve the final total score. Raters are asked to make provisional assessments after each role play, and at the end of the interaction to record a final definitive assessment.

For *writing* there are five assessment categories, as follows: *Overall task fulfilment* (weighted in the same way as the equivalent overall category in the speaking test); *Appropriateness of language*; *Comprehension of stimulus*; *Control of linguistic features (grammar and cohesion)*; and *Control of linguistic features (spelling, punctuation)*.

3.5 Writing and trialing of pilot versions of test materials

The development of materials for the Speaking and Writing sub-tests involved securing the cooperation of expert informants. Informants were identified within each of the professions involved in the test. These were usually educators within the profession. For the Speaking sub-test materials, each was asked to identify topics commonly occurring in consultations which could form the basis of role play materials. The topics should not involve or appear to involve anything other than basic professional knowledge, as this was to be assessed exclusively in the subsequent pencil and paper and clinical tests. Nor should the materials involve too complex a

task for the interlocutor, who would obviously not have any medical expertise, and could not be expected to invent plausible symptoms, for example. Ideas for role play situations were subsequently written up as role play cards, and shown again to the informant for further comment.[9] A second informant from within the profession was then shown the role cards and asked to comment. A final revised version of the role cards was subsequently prepared for trialing (cf Appendix C for sample materials).

A similar process took place for the *writing* materials, except that this time the informants were commissioned to write the case notes as the stimulus for the letter of referral *and* a typical letter that might result, to be used as a reference point by the raters. Again, the materials were shown to another informant for comment prior to final revision. Sample materials can be found in Appendix D.

Reading test materials were drawn from suitable professional journals, as well as more general sources. Topics were chosen so as to be accessible to the range of professionals taking the test; multiple choice questions were then written.

For the first part of the Listening sub-test, speakers familiar with the task of conducting professional upgrading seminars for health professionals were sought. A studio recording of a talk on a suitable topic was made and transcribed, then edited, and test questions written. A solution to the problem of producing authentic listening material without breaking confidentiality was found for the consultation. It was discovered that a 'simulated patient' service was available, supplying actors who had learned the symptoms of an actual patient with a particular condition and who would then 'present' as the patient in a simulated consultation. (This service is used in the training of health professionals in communication skills in interaction with patients.) Actors from the service were hired and paired with medical practitioners who were instructed to take medical histories from them in the usual way. The interaction, which was unscripted, achieved remarkable authenticity as the medical practitioners said that they found the actors and the whole

[9]In the development of subsequent versions of the test, the writing up was done by the informant and checked by the test developer and another expert informant.

simulation credible and natural and were then able to fall naturally into the familiar role of history taker.

The setting of an appropriate listening task proved more problematic, however. The test format requires candidates to take case history notes as they listen. At first, candidates were allowed freedom over the form of these notes, which were marked for content. However, the absence in the candidate's response of a detail of the history deemed relevant and required in the marking scheme was felt to be ambiguous. It was not clear whether it signified that the detail had not been heard or understood, or that it had not been judged worth noting by the candidate. In fact, expert informants explained that in reality note taking in case histories varies considerable from practitioner to practitioner: some take detailed notes, others almost none. In any case, note taking will be organised around diagnostic considerations; if the diagnosis is clear then a brief note may suffice. But in order to get a *scorable* listening performance, note taking had to be constrained in a potentially quite inauthentic way: candidates were instructed that notes were required under specific content headings.

The integration of expert professional knowledge and observable language behaviour in this instance goes to the heart of the difficulty of the work sample tradition of performance testing. Authentic performances cannot easily be replicated in the test setting; writing on this subject in the language testing literature often appear rather complacent on this point. Appearances have often disguised deeper issues of validity.

4. Limits to simulation as basis of test content

At the beginning of this paper, it was stated that replication of the criterion in the test could be done in relation to each of: task stimulus; task demand; task processing; and the criteria by which successful performance of the task is evaluated. We will now consider the limits on the replicability of each of these in the light of the above discussion.

Task stimulus

There is an inevitable compromise across and within professions. Even though the research established commonalities across the professions, for example in the area of case history taking or

assessment, the form that this takes inevitably differs from profession to profession. The medical case history used as the basis of the listening test may not be exactly like those found in other professions. Similarly within professions, there is considerable difference between, for example, speech pathology with children and adults.

Further, it is possible that simulations may distort the reality of professional communication situations when implemented under test conditions. In the OET, materials are written by health educators in conjunction with test developers, so that they have some claim to authenticity at the design stage. Lumley and Brown (1996) have researched the reactions of health professionals to the realism of the role play communication elicited under test conditions. They found some grounds for optimism as most interactions were deemed to be relatively authentic, but there were criticisms of some materials; considerable ongoing care and monitoring on this point seem to be justified.

Task demand

It is difficult to simulate task demands. We showed above that in the Listening sub-test, the authentic practice of note taking needed to be modified in the light of test requirements. That is, there was a need to constrain the form of the candidate's response in a way which is not faithful to criterion behaviour.

Task processing

Again, this aspect of the criterion is very difficult to simulate. First, communication in medical contexts requires an integration of medical knowledge and language skill. This is clear from the note taking example just mentioned, where for most practitioners notes are organised around diagnostic frameworks; similarly, discourse organisation in case history depends crucially on diagnostic hunches. Simulations of doctor-patient communication using role play in first language medical education contexts result in performances which are very different from those in second language tests such as the OET (see McNamara, 1996: 82-84 for details). Further, the psychological conditions of communication are inevitably rather different in test and stressful clinical situations

Success criteria

This is perhaps the weakest area of simulation, and arguably the most crucial. I have proposed elsewhere (McNamara, 1996: 43-45) a distinction between strong and weak performance tests, depending on the degree to which real world criteria are used in evaluating performances. A further issue here is to establish by what standards performances are in fact evaluated in reality - we know little about this, as I argue in a recent paper with Sally Jacoby (Jacoby and McNamara, forthcoming) on the naturally occurring contextualized assessment of communication that goes on in the course of everyday interaction at work (indigenous assessment). A related question is whether the same standards do or should apply for native and non-native speakers in professional settings - on the face of it, they should be identical, given that the occupational roles and responsibilities are identical.

Given the difficulties presented in this discussion, we may ask why we should bother to simulate at all, given its inevitable superficiality? There are two main arguments: the face validity one, that it is important for the acceptability of a test such as the OET which operates in a socially and politically high stakes context; and the arguments for the positive washback of the test, that is its assumed beneficial impact on teaching leading up to it.

5. Conclusion

In this paper, the stages in the development of work-sample performance tests have been outlined, and the carrying out of such a development project has been described in some detail. The account shows some of the difficulties encountered at the practical level in the development of a work sample test, and the kinds of decisions that are forced on the test developer, together with their implications for the validity of the resulting test. In particular, the detail of the stage of content specification raises issues about the validity of this and other work sample tests on a number of grounds, including the difficulty of simulating authentic communicative process under test conditions, and the inseparability within authentic performance of professional knowledge and language behaviour. Work sample tests such as the OET are strong on face validity, but any claim that the OET fulfils its brief of a test of 'the ability of candidates to communicate effectively in the workplace' (Alderson *et al.*, 1986: 3) must be interpreted with caution. This is

not to say that the OET is a bad test; in fact it is a carefully constructed test and one that is administered in an entirely professional way. Rather, it illustrates the real difficulty of language performance testing in work-related contexts.

References

Alderson J C, Candlin C N, Clapham C M, Martin D J, Weir C J (1986) Language proficiency testing for migrant professionals: new directions for the Occupational English Test. A report submitted to the Council on Overseas Professional Qualifications. Lancaster: Institute for English Language Education, University of Lancaster.

Byrne, P.S. and B.E.L. Long (1976) Doctors talking to patients. Her Majesty's Stationery Office, London.

Candlin, C.N., Leather J H, Bruton C J (1974) English language skills for overseas doctors and medical staff. Work in progress, Reports I-IV. Lancaster: University of Lancaster, Department of Linguistics and Modern English Language.

Coulthard M, Ashby M (1975) Talking with the doctor, 1. Journal of Communication 25, 3: 140-47.

Davies, A. (1977) The construction of language tests. In Allen J P B, Davies A (eds) Testing and experimental methods. The Edinburgh Course in Applied Linguistics vol 4. Oxford: Oxford University Press, pp 38-104.

Davies, A. (1984) Validating three tests of English language proficiency. Language Testing 1, 1: 50-69.

Emmett A (1985) The Associated Examining Board's Test in English for Educational Purposes (TEEP). In Hauptman P C, LeBlanc R, Wesche R (eds) Second language performance testing. Ottawa: Ottawa University Press, pp 131-51.

Jacoby, S. and T.F. McNamara (forthcoming) Locating competence. English for Specific Purposes 17,4

Lumley, T. and A. Brown (1996) Specific-purpose language performance tests: task and interaction. In G. Wigglesworth and C. Elder (eds) The language testing cycle: from inception to washback. Australian Review of Applied Linguistics, Series S, Number 13: 105-136.

Martin J R (1984) Language, register and gender. In ECT418 Language Studies: Children Writing. Victoria: Deakin University.

Mawer G (1991) Language audits and industry restructuring. Sydney: National Centre for English Language Teaching and Research, Macquarie University.

McNamara T F 1987 Assessing the language proficiency of health professionals. Recommendations for reform of the Occupational English Test. A Report submitted to the Council on Overseas Professional Qualifications. Parkville, Victoria: University of Melbourne Department of Russian and Language Studies.

McNamara T F 1988 The development of an English as a Second Language speaking test for health professionals. Part One of a Report to the Council on Overseas Professional Qualifications on a consultancy to develop the Occupational English Test. Parkville, Victoria: University of Melbourne Department of Russian and Language Studies.

McNamara, T.F. (1989) ESP testing: general and particular. In C.N. Candlin and T.F. McNamara (eds) Language, learning and community. Sydney: National Centre for Eglish Language Teaching and Research, Macquarie University, 125-142.

McNamara, T.F. (1996) Measuring second language performance. London and New York: Addison Wesley Longman.

Messick, S. (1989) Validity. In R.L. Linn (ed.) Educational measurement. Third edition. New York: Macmillan, 13-104.

Messick, S. (1994) The interplay of evidence and consequences in the validation of performance assessments. Educational Researcher 23, 2: 13-23.

Morgan J (1994) Literacy and Productivity. Paper presented at Language Expo Australia, Sydney, July

Paltridge B (1994) Genre analysis and the identification of textual boundaries. Applied Linguistics 15, 3: 288-99.

Powers D (1986) Academic demands related to listening skills. Language Testing 3, 1: 1-38.

Skopek L (1979) Doctor-patient conversation: a way of analyzing its linguistic problems. Semiotica 28, 3/4: 301-11.

Swales J (1990) Genre analysis. Cambridge: Cambridge University Press.

Weir C J (1983a) Identifying the language problems of overseas students in tertiary education in the United Kingdom. University of London unpublished PhD thesis.

Weir C J (1983b) The Associated Examining Board's Test in English for Academic Purposes: an exercise in content validation. In Hughes A, Porter D (eds) Current developments in language testing. London: Academic Press.

Wilds C P 1975 The oral interview test. In Jones R L, Spolsky B (eds) Testing language proficiency. Arlington, VA: Center for Applied Linguistics, pp 29-44.

Appendix A: Extract from scoring grid - speaking sub-test

OVERALL COMMUNICATIVE EFFECTIVENESS

Near-native flexibility & range _ | _ | _ | | _ | _ | _ Limited

INTELLIGIBILITY

Intelligible _ | _ | _ | | _ | _ | _ Unintelligible

FLUENCY

Even _ | _ | _ | | _ | _ | _ Uneven

COMPREHENSION

Complete _ | _ | _ | | _ | _ | _ Incomplete

APPROPRIATENESS OF LANGUAGE

Appropriate _ | _ | _ | | _ | _ | _ Inappropriate

RESOURCES OF GRAMMAR AND EXPRESSION

Rich, flexible _ | _ | _ | | _ | _ | _ Limited

Appendix B: Extract from scoring grid - writing sub-test

OVERALL TASK FULFILMENT

Completely satisfactory _ | _ | _ | | _ | _ | _ Unsatisfactory

APPROPRIATENESS OF LANGUAGE

Appropriate _ | _ | _ | | _ | _ | _ Inappropriate

COMPREHENSION OF STIMULUS

Complete _ | _ | _ | | _ | _ | _ Incomplete

CONTROL OF LINGUISTIC FEATURES (GRAMMAR AND COHESION)

Complete _ | _ | _ | | _ | _ | _ Incomplete

CONTROL OF PRESENTATION FEATURES (SPELLING, PUNCTUATION)

Complete _ | _ | _ | | _ | _ | _ Incomplete

Appendix C Sample materials - Speaking sub-test (physiotherapists)

CANDIDATE'S CARD

SETTING: Hospital clinic

PATIENT: An elderly person who is recovering from a stroke (CVA). The patient is making slow progress in learning to walk again.

TASK: Talk to the patient about the following pieces of equipment

- a wheelchair
- a walking frame
- a walking stick.

Explain the advantages and disadvantages of each one.

You would like the patient to be as independent in his/her movements as possible. You feel the frame is not appropriate.

You want the patient to have a stick. You do not want the patient to have a wheelchair at this stage.

ROLE PLAYER'S CARD

SETTING: Hospital clinic

PATIENT: You are an elderly person who is recovering from a stroke. You feel you are making painfully slow progress, and don't really expect to be able to walk again.

You feel you should be allowed to have a wheelchair.

TASK: Ask the physiotherapist when you will be given a wheelchair.

Insist on your need for this equipment. Explain that you feel that the painful exercises you are doing at the moment are pointless, and that you are pessimistic about your chances of making real progress.

Be difficult!

Appendix D Sample materials - Writing sub-test

OCCUPATIONAL ENGLISH TEST - WRITING TEST (MEDICAL PRACTITIONERS)

Time allowed: 40 minutes

Mrs Lyons is a patient in your general practice. Read the case notes below and complete the writing task that follows.

CASE NOTES *Mrs Harriet Lyons* 84 yo woman.

14/5/88

PH:
- osteoarthritis ® hip → THR 1985.
- hypertension x 20 yrs.
- Type II diabetes x 15 yrs.
- recurrent UTIs
- dementia x 10 yrs.

Medications:
- Daonil 5 mg bd
- Aldomet 500 mg bd
- Indocid 25 mg tds

Brought in by daughter, with whom she lives.
Increasingly difficult to cope with her.
- urinary incontinence for last week. ?dysuria
- abdominal pain.
- No fevers/sweats/loin pain.
- More confused than usual. Refusing to eat.
- No vomiting, diarrhoea.

O/E: Afebrile. Confused.
Mild suprapubic tenderness.
Urine: protein +++ RBC +++ glucose 1/2%.

Assessment: Worsening mental state 2° to UTI.
As MSU impossible to obtain,
R with Amoxil 500 mg tds x 7 days.

21/5/88 No more incontinence. Confusion improved.

12/6/88 Found wandering in the street by neighbours. Becoming increasingly vague. No other specific symptoms. Daughter very tearful. Reassured.

4/7/88 Found lying next to bed by daughter. Tripped

cont...

	over rug on way to toilet. Incontinent. Behaviour becoming more difficult lately; emotional outbursts, refusing to cooperate.
	Unsteady gait recently.
O/E:	Confused. BP 140/75 lying 110/60 standing. Bruise on ® hip. Movements good. No other injuries noted.
Assessment:	Postural hypertension 2° to Aldomet.
For:	↓ to 250 mg bd.
21/7/88	Gait has improved, but mental state continuing to be a problem. Daughter feels that she 'just can't cope any more' without outside help. Thinks that 'a nursing home might be best for everyone' and requests specialist opinion.
For:	Refer to Dr Chalming (geriatrician) re improved medical management and/or placement.

WRITING TASK

Using the information in the case notes, write a letter of referral to Dr Chalming. The main part of the letter should be 12-15 lines long.

Do not use note form in the letter; expand the case notes where relevant into full sentences.