Language Proficiency testing in health settings
Report No. 4

This report is part of a series of documents:

Report 1 Health communication between non-English speaking patients and bilingual staff within our health services. ISBN 1 875 909 89 3
Report 2 Matching non-English speaking patients and bilingual staff within wards or units. ISBN 1 875 909 90 7
Report 3 Bilingual staff in mainstream healthcare: Policy development for NSW Health Services. ISBN 1 875 909 91 5
Report 4 Language proficiency testing in health settings. ISBN 1 875 909 92 3

Communicating across language and culture in the hospital system series

Sponsored by: Health Research Foundation Sydney South West.
Faculty of Health, University of Western Sydney (Macarthur).
Multicultural Service Enhancement Program, New South Wales Health.

Additional funding and/or support from:
South Eastern Sydney Area Health Service
Language Testing Research Centre, University of Melbourne.
Faculty of Education and Languages, University of Western Sydney (Macarthur).

Reference this report as:
Elisabeth Grove, Annie Brown. Language proficiency testing in health settings. Report No.4. in the *Communicating across language and culture in the hospital system* series. South Western Sydney Area Multicultural Health Service and the South Western Sydney Centre for Applied Nursing Research.

ISBN 1 875 909 92 3

October 2000.

For additional copies please contact:

Centre for Applied Nursing Research
South Western Sydney Area Health Service
Liverpool Health Service, Locked Bag 7103,
Liverpool BC NSW 1871
Telephone +61 2 9828 6537 Fax +61 2 9828 6519

TABLE OF CONTENTS

ACKNOWLEDGEMENTS 1
INTRODUCTION 2
GENERAL ISSUES IN TESTING LANGUAGE FOR SPECIFIC PURPOSES 3
FEASIBILITY STUDY 5
The Bilingual Health Communication Model *5*
Levels of Fluency *6*
Implications for test design 6
Contexts of Language Use *6*
Implications for test design 7
Potential Test Candidates *8*
Test Purpose *8*
General Recommendations for the Test *8*
The Recommended Test Model *9*
Language Background and Self-Assessment Questionnaire 9
Rationale 9
A performance test of occupation-specific oral skills 10
Oral test model 10
TEST DEVELOPMENT 10
Production of Pilot Assessment Instruments *10*
Language Background and Self-Assessment Questionnaire *11*
Oral Test Content and Procedures *11*
Summary of Tasks *12*
Task 1: Giving directions 12
Task 2: Explaining procedures 12
Task 3: History-taking 12
Task 4: Patient Education (nursing and allied health staff only) 12
Task 4: Informed consent (medical staff only) 13
Task 5: Health terminology 13
Pilot Assessment Criteria *13*
Summary of draft band levels and assessment criteria 13
TRIALLING THE PILOT ASSESSMENT INSTRUMENTS 14
Range of Pilot Candidates *14*
Selection of Pilot Test Participants *14*
Procedure to Recruit Participants *15*
Recommendations for Future Test Administration *16*
Interviewer Training Workshop *16*
Results of Pilot Tests *17*
Candidate profile 17
Candidate profile by language 17
Test Results *19*
Interviewer assessments 19
Assessment procedure 19
Summary of results 20
Criteria 21
Task difficulty 21
Standard-Setting Workshop *22*
Recruitment of health professional informants 22
Assessment procedure 22
Results of assessments 23
Comments 24
Self-Assessment *25*
Results 25
RECOMMENDATIONS AND ISSUES FOR CONSIDERATION 26
REFERENCES 29
APPENDIX 1: LANGUAGE BACKGROUND QUESTIONNNAIRE: MEDICAL, NURSING AND ALLIED HEALTH STAFF 30
APPENDIX 2: BILINGUAL HEALTH COMMUNICATION SKILLS TEST ASSESSMENT SCALE 36
APPENDIX 3: BILINGUAL HEALTH COMMUNICATION SKILLS TEST 38

List of Tables

Table 1: Situations of Language Use at Work by Bilingual Staff (excluding Ethnic Health Staff/Interpreters) ... 7

Acknowledgements

Thanks are due to many people, without whose efforts the project could not have gone ahead:

Associate Professor Anna Whelan, for her invaluable work on all aspects and stages of test development.

Associate Professor Cathie Elder, former director of the Language Testing Research Centre - for setting up the testing project and securing the Melbourne University Collaborative Research Grant which enabled further work to be undertaken

Members of the Multicultural Health Research Team – Clair Matthews, Cathy Noble and Professor Maree Johnson – for their advice and critical commentary on the feasibility study and draft versions of the test.

LTRC staff - Noriko Iwashita, Ruta Kanepe, Sally O'Hagan and Erich Round – for their help with the preparation of test materials and handling of test data.

Staff of the South Western Centre for Applied Nursing Research - Soufiane Boufous, Venita Devi, Elizabeth Halcomb, Rachel Langdon and Ritin Fernandez – for help with all aspects of test organisation.

The bilingual health professionals who participated in the pilot test – nurses, doctors, allied health, multicultural health and support staff.

Unit managers and other area health staff who assisted with the practicalities of test administration.

The interviewers – Giovanna Ng, Wai Hung Lam, Andy Nguyen and Ninh Nguyen - for dealing so patiently with the demands of the pilot test administration and for providing feedback on test content and procedures.

The occupational experts who participated in the standard-setting exercise – Dr Bee Hong Lo, Grace Wong, Dr Ngo and Bach Tuyet Nguyen.

Professor Loretta Giorcelli of the Division of Languages, at the University of Western Sydney, Department of Education and Languages, for identifying suitable interpreters to act as interviewer assessors.

Sam Choucair of SESAHS, for supporting the testing project with a significant contribution to the funds for test development.

SESAHS staff - Jo Travaglia especially, for their help in contacting potential candidates and encouraging them to participate in the test.

Irene Liem, for her work on the Cantonese test samples, and Dr Paul Ghaie, for advice on the tasks for medical candidates.

Introduction

Reports 1 and 2 have amply demonstrated the importance of shared language for good communication in health care and unravelled the complexities of utilising the linguistic resources of bilingual staff within the health system. In particular, the Multicultural Health Research Team has emphasised the need to encourage bilingual staff to use their skills in communicating with patients *within the scope of their language proficiency*. This emerged as a crucial issue in the 1996/7 Language Audit of SWSAHS employees, which found that 30% of 'mainstream' staff were bilingual or multilingual. SWSAHS recognised a major source of language (and cultural) skills which might be used to benefit patient care, and which should be formally acknowledged within the Health Service. However, it could not be assumed that those who may be described as bilinguals, or 'background speakers' of community languages, possessed comparable levels of language ability or could perform all of their work duties in the LOTE (Language Other Than English). Factors such as level of education in the other language, length of time in Australia and age of immigration have a direct impact on the individual's ability to use the language for various communicative purposes. Among the staff surveyed, the Language Audit detected a range of perceived levels of proficiency and language use, from migrants trained professionally in their first language, to bilinguals who used their language at home, to second- or third-generation migrants who mainly used English but were able to speak, or at least understand, another language.

With such diverse levels of LOTE proficiency among staff, it was recognised that encouraging the use of their language skills in the direct care of patients must be accompanied by controls: to protect staff from coercion to use their language inappropriately, patients from staff who are not linguistically competent, and the Area Health Service, from medico-legal problems. As indicated in Report 2, the Multicultural Health Research Team is of the firm view that all bilingual health staff wishing to use their language skills in communication with patients should have their proficiency assessed appropriately. Currently, however, there is no accepted method for measuring the proficiency of bilingual staff. In response to this need, SWSAHS sought collaboration in 1998 with the Language Testing Research Centre (LTRC), The University of Melbourne, to investigate the feasibility of developing a formal language assessment instrument of oral skills in three community languages (Arabic, Spanish and Vietnamese).

The instrument was required to be valid and reliable, capable of assessing both simple and more complex health language competence, and easy and inexpensive to administer. The LTRC was well placed to undertake the task, with considerable experience of developing tests of language for specific purposes (LSP) for use in the public sphere.[1] The initial brief was later revised to

[1] Relevant tests developed by the LTRC include health-related language tests such as the Occupational English Test for Health Professionals (OET), developed for the National Office of Overseas Skills Recognition (NOOSR) and the Health Sciences Communication Skills Test, developed for the University of Melbourne. Specific-purpose LOTE tests include the Japanese Test for Tour Guides, developed for Tourism

include two languages, Cantonese and Vietnamese, with the intention of later extending the test to other important community languages.

The next section of Report 4 outlines the recommendations of the Feasibility Study and the rationale for the test model eventually selected. A number of important issues involved in LSP testing and their implications for the pilot test are then discussed, followed by an account of the process of test development. Test content, the selection of candidates and test administration procedures are described next, and the results of pilot testing and the preliminary validation presented. The report ends with recommendations for further test validation and extension of the test to additional languages.

General Issues in Testing Language for Specific Purposes

Testing language for specific occupational purposes is undertaken in order to determine whether or not individuals possess the linguistic ability to carry out particular work tasks in the target language. In the past, relatively indirect test tasks (e.g. of grammatical knowledge) were used and the results taken to predict the language user's ability to function in the real life context. In recent years, however, there has been an increasing trend to develop more direct tests, which seek to simulate the work environment, the context of language use. Such tests are commonly known as performance tests.

The oral interview component of the pilot Bilingual Health Proficiency Test is a performance test, that is, a test in which candidates are required to perform a simulation of the actual target tasks. Such tests are well suited to situations where the target situation can be clearly delineated and described. As Jones (1985) puts it, *'It is impossible for a language test to predict task-oriented proficiency unless it includes or approximates actual samples of the tasks.'* In the development of the test tasks it was therefore important to involve 'industry' representatives (i.e. health professionals) in the identification of the types of interactions which are central to the work of communicating with patients. However, because the range of professions (nursing, medical and allied health staff) and the variety of specialisations within them are so great, it was also necessary to try to narrow the range of skills and tasks in order to produce a manageable and practicable test instrument.

It is obviously not possible to *test* that LOTE-users have the requisite language knowledge for *all* domains[2]. A test can only *sample* from the relevant domains, and inferences are then made about the test-taker's ability to perform in other domains**.** However, legitimate generalisation from one performance (test) task to another (non-test) task can only be made if they present equivalent linguistic demands. Sampling involves choosing a selection of tasks from the domain of language use under test. The adequacy of the sample is essential to establishing the *content validity* of the test, (that is, whether the tasks are representative of the

Training Australia, and Teacher Tests in French, Italian, Indonesian and Japanese developed for The National Languages and Literacy Institute of Australia (NLLIA).

[2] The term 'domain' is used to refer to language use associated with a specific area of activity, particularly for occupational purposes.

domain). However, as Davies et al (1994:174) point out, '*the inherent variability of language makes it difficult to ensure that the sample selected is adequate'*. In order to assess the adequacy of task sampling for this test, it will be necessary to investigate other aspects of test validity, such as its *predictive validity* (that is, how well the test results predict the individual's ability to perform in the real work context). Follow-up case studies of staff members who have sat the test would be the most direct means of investigating this.

Another important issue that arises in assessing a candidate's ability to use language for work communication, is whether or not the language skills demonstrated in the performance of a given task can (or should) be distinguished from the other aspects involved in successful communication, '*the general non-language based communication skills and traits which also affect the listener's evaluation of the quality of the performance*' (Brown 1992:39). In some LSP tests, the Occupational English test for Health Professionals (OET), for instance, there is an exclusive focus on the linguistic features of candidate performance. In the OET, the role play, a simulated professional communication task between health professional (candidate) and 'patient' (interviewer), is essentially a stimulus to elicit a sample of language for assessment. The task therefore has no more than *face validity* (which is concerned with the appearance of validity rather than with the underlying construct of language ability being measured by the test). The decision to restrict attention to linguistic features (such as comprehension, grammar, intelligibility, resources of grammar and vocabulary) was made in order to avoid the situation of language-trained assessors being required to assess the professional knowledge and skills of the candidates.

McNamara (1990) has called this type of test a 'weak' performance test, as distinct from a 'strong' test of performance (such as a test of clinical medical knowledge). In a 'strong' language test, he argues, *'language ability will be only one of the many criteria used in assessing performance. Performance will primarily be judged on real-world criteria, that is, the fulfilment of the task set.'* This strong-weak distinction has been criticised in recent years because of the practical and theoretical difficulty of establishing and maintaining (Davies, 1995; Douglas, 2000). In Davies' view, all specific-purpose tests are '*more or less strong*'. This is because

> *making the right language choice cannot be judged in terms of language alone; of necessity recourse must be made to context. In the one case [strong tests] knowledge needs language to encode it; in the other [weak tests] language needs knowledge or content to give it meaning.* (p.11)

The ongoing debate among language testers is highly pertinent to the issues that will need to be resolved in relation to the assessment of bilingual health professionals in this project. It is also central to the perceptions of potential test-takers to the test, and their willingness volunteer to be assessed. This was evident in the anecdotal evidence from bilingual doctors cited in Report 2 of this Series (p.18).

> *If you are testing communication skills, that is directly linked to how good a doctor they are, and is part of being a doctor. It will be difficult to test it. It is going to be important to separate testing language skills from communication skills. How reliable will the assessment be?*

It will be necessary to establish as clearly as possible that both the intention (and the reality) of the test is to assess *the linguistic ability of staff to perform their work duties in the LOTE*, not their professional skills as communicators. However, as already indicated, this distinction may be very difficult to maintain and will require further investigation. (See the Section on the standard-setting workshop for further details)

Feasibility Study

The feasibility study (Grove and Brown, 1999) which established the foundations of this collaborative testing project investigated the theoretical and practical issues raised by the SWSAHS brief, outlined the implications for test development, reviewed a number of relevant existing tests, and recommended the development of a new test of oral communication to serve the range of needs identified. A range of test options was suggested and an estimate of the comparative costs and benefits of each provided.[3]

This section of the report summarises the main needs identified and the rationale for the test model finally selected.

The Bilingual Health Communication Model

The model of health communication developed by the project team (Johnson, Noble, Matthews, Aguilar, 1997, 1998, 1999) was used as a basic framework for test development, and as the basis for the construct of language ability which the test set out to measure. The project team identified a range of communication situations and a model of bilingual language use, which they have called The Bilingual Health Communication Model: A Matrix of Fluency and Context of Interaction. The model is presented as two intersecting continua, linking the types of communicative functions, the tasks and the levels of complexity of the communicative skills used in patient-health worker interactions, based on how bilingual staff regarded their fluency[4] in the LOTE and the context of situations in which they used it (Johnson et al 1999). The matrix indicates that the majority of bilingual health workers use their language skills within the area of simple

[3] It was made explicit in the feasibility study that, while the development of suitable instruments for assessing the LOTE skills of bilingual health workers was feasible, the limited funds available for test development, administration and maintenance imposed significant financial constraints. While there were sufficient funds for basic test design and development, interviewer training and analysis of pilot test results, additional funds were needed for test validation. In November 1999, the combined LTRC-SWSAHS/SESAHS team succeeded in gaining a further $7,500 from the University of Melbourne Collaborative Grant Scheme to fund a preliminary validation study.

[4] In linguistic usage, the term 'fluency' generally refers to 'the features which give speech the qualities of being natural and normal, including native-like use of pausing, rhythm, intonation, stress.' (Richards et al, 1993:141) For that reason, throughout the report, we have used the terms 'proficiency' and competence' to refer to the 'degree of skill with which a person can use a language' (ibid:205)

communication, with smaller numbers regarding themselves as complex verbalisers who also used their LOTE in the complex medico-legal domain.

Levels of Fluency

Three distinct groups were identified on the fluency continuum:

1. No fluency but cultural awareness and understanding.
2. Social fluency or language fluency in the home situation.
3. Sophisticated fluency (able to articulate and negotiate complex interactions across social and professional domains). (Johnson et al 1997:38)

Implications for test design

The most noteworthy feature of this continuum in terms of test design is that, while it identified three distinct groups, for assessment purposes, there were really only *two levels of linguistic competence*: social fluency and sophisticated fluency. The first, 'no fluency but cultural understanding', indicates that individuals in this group possess no effective ability at all to use the language. The second, social fluency, is extremely broad, and, as the researchers have pointed out, was the one in which there was the greatest variation in perceived fluency levels. This in itself implied the need for a means of assessment capable of making *distinctions* in terms of what individuals are able to do in the LOTE. The potential range of competence within this group thus needed to be more precisely defined, which required the use of a more extensive scale. The third group, sophisticated fluency, was also thought to conceal important differences which the test needed to take into account: a high level of general proficiency does not necessarily entail possession of linguistic (especially vocabulary) skills adequate to interacting or interpreting in domains for which the individuals are not trained. For the purposes of language assessment, a two- or even three-point scale is inadequate to allow *different levels of language proficiency to be linked to the ability to perform tasks over a range of difficulty or complexity*.

Another problem concerned the source of information on which the fluency continuum was developed, namely, the health workers' own perceptions of their language competence. Issues relating to the use of self-assessment will be dealt with further in 'Test options'. At this point, it is enough to note that a) it cannot be used as a reliable measure of proficiency, particularly for the purpose of public accreditation of competence, and thus, b) it cannot form the basis of an accurate scale against which performance can be mapped. (Note: the draft assessment scale used for the pilot test was therefore designed to cover *four levels* of proficiency.)

Contexts of Language Use

The context-of-use continuum identified a range of contexts and situations of LOTE use, categorising them into two broad categories, involving 'simple' and 'complex' language, as follows on the next page:

Table 1: Situations of Language Use at Work by Bilingual Staff (excluding Ethnic Health Staff/Interpreters) (adapted from Johnson et al 1997:33)

Simple Language	% response
Simple language (basic social exchanges)	17.6
Giving directions	13.1
Registering/booking	4.0
When patients are upset	9.1
Identification of problems and giving explanation	11.3
Complex health language	
Taking medical history/assessing medical condition	8.7
Explanation/consent for/ procedure	8.4
Consent for release of information	2.0
Written consent	1.5
Ongoing treatment	9.1
Education	8.3
Counselling/therapy	4.5
Other situations	2.2

Implications for test design

From a linguistic point of view, the simple-complex health language dichotomy proposed by the research project team in the 1997 report was problematic. For the purpose of test design, there were several issues to be considered:

- Within both categories, there may be situations and tasks that require *differing levels of linguistic competence*, differences which cannot be accounted for by a simple-complex dichotomy. For instance, an individual may be able to perform all or some tasks within both categories. In order to decide what to do with such information, in terms of accrediting individuals to use their LOTE skills, it would in our view be necessary to discriminate more precisely. (Note: This need for greater specificity and discrimination was later incorporated into the description of draft band levels and to the graded sequence of tasks)

- The percentage frequencies of these 'communication situations' in patient-health worker interactions do not necessarily relate to the *level* of language proficiency required to perform them. They may also reflect the more general *distribution* of such communication situations in the daily routine of health workers, monolingual and bilingual alike.

- Assumptions made about the relative *complexity* of tasks may have more to do with the health professionals' perceptions of the ease and familiarity of carrying out the procedures than with the linguistic skills and resources needed to communicate. For instance, resolving a problem for a patient may involve the use of quite complex grammatical structures and sensitive choice

of vocabulary, but require less specialised knowledge than is needed for explaining a technical procedure. (Note: similar observations by the health professionals themselves are quoted in Report 2)

It was evident that more investigation would be needed by the project team and the test developers of what constitutes linguistic simplicity or complexity in patient-health worker interactions. In terms of language assessment, the linguistic requirements of these tasks needed to be established so that the tasks could be mapped onto levels of performance on the test. This would also require later validation and monitoring (e.g. via observation) of workers who have been assessed at particular levels so that the accuracy of the levels can be verified. It would also be necessary to investigate the issue more fully in the process of test development through a more detailed analysis of

1. The types of linguistic skills used in health workers' routine tasks.
2. The language levels needed to undertake specific tasks.
3. The type of content which would be appropriate to include in the assessment tasks.

Potential Test Candidates

It was assumed that bilingual staff whose professional training has been undertaken in their first language would not need to be formally tested. The test is therefore concerned to assess the proficiency of staff who have not used the LOTE as the primary language of their professional work, but who have demonstrated (or claimed) an adequate level of general proficiency in their self-assessment of proficiency.

Test Purpose

The assessment instruments would be required to serve two purposes:

- to encourage staff to use their language skills appropriately, particularly in routine social communication with patients, but also
- to provide appropriate control of the language use of the relatively small numbers of staff who might be capable of carrying out all or most of their duties in the LOTE.

It was also recognised that these two purposes were potentially conflicting: to encourage staff to use their language skills, while also imposing restrictions on that use via a formal test of higher level skills, might be perceived as contradictory or confusing.

General Recommendations for the Test

LTRC proposed the development of a new test of occupational language skills to assess the ability of health professional LOTE-users to function professionally in the language. In order to satisfy the dual purpose of the testing instrument, it was recommended that two distinct, but linked, assessment procedures should be developed.

For staff to undertake all of their own duties in the LOTE, certain specialist language skills are needed, skills that even a native-speaker cannot be assumed to be competent in without training. These include control of semi-technical and technical terminology and procedural language (e.g. obtaining a patient's informed consent to a procedure or explaining a procedure).

After consultation with SWSAHS, it was agreed that two test instruments of oral (spoken) communication should be developed: one of the 'simple social language' used in routine daily interactions with patients, the other to assess linguistic competence to deal with the transmission of more complex 'technical' information and to handle medico-legal procedures (such as obtaining patient consent).

The Recommended Test Model

Language Background and Self-Assessment Questionnaire

In order to build up a database of bilingual staff resources, it was decided that background information on bilingual workers should be gathered by means of a questionnaire, to include details of educational level, professional position, length of exposure to the LOTE and general contexts of use. It was recommended that all staff who indicated in the questionnaire that they are native speakers not using the language regularly to perform their work duties should be assessed for general competence in the language, via self-assessment.

Rationale

- As pointed out in the previous section, self-assessment cannot be used as a reliable proficiency measure for the purpose of accreditation of skills. However, self-assessment was proposed as the best way to ensure that adequate numbers of staff would self-select to use their LOTE in *social and support contexts*. It was also essential to ensure that, especially for crucial communication with medico-legal implications, a valid and reliable measure would be developed.
- A prime concern of the project was to encourage as many staff as possible to self-select to use their LOTE, and LTRC was aware that the requirement that staff (albeit a sample) undertake a test on top of the questionnaire might be seen as unnecessarily de-motivating. However, we believed that if staff were prepared to put themselves forward to use their LOTE for professional purposes they should also be prepared to undertake a simple assessment to verify the information they gave about their language skills. An external measure is the only way the reliability of the self-assessments can be verified. As an alternative to independent assessments, the effectiveness[5] of the instrument in this case may be monitored through a follow-up of the experiences of staff self-selecting into the LOTE-user category, either via on-the-job observation or review of their own performance (through interview or survey).

[5] 'Effectiveness' is a holistic concept which includes both reliability and validity.

It should be noted here that the issue of reliability or effectiveness has cost implications, as it would require either a follow-up study or a *concurrent validation* (comparison of self-assessments with independent ones). Present funds have not permitted more than a preliminary investigation of this issue.

A performance test of occupation-specific oral skills

In order to obtain the broadest and most reliable information on candidates' abilities, it was proposed that the specialist test consist of a *performance* component and a *written* component. However, this relatively expensive option was later re-conceived by adding a test of health-related vocabulary to the end of the interview. (For details, see section on Oral Test Content.)

Oral test model

The test would be designed as an interview, consisting of five tasks sampling the occupational domain, in order to assess LOTE-users' task-based procedural language skills. The test tasks were to be selected on the basis of their frequency or importance (such as obtaining a consent and explaining a procedure), along the lines described by the Bilingual Health Communication Model. On the basis of the literature search and consultation with the research team, the test designers assumed that the same generic skills are relevant to a range of the communication tasks performed by both nursing and medical staff, and so decided to provide common versions of the test tasks concerned with social exchanges, explaining simple procedures and history-taking. However, it was decided that there should be alternative versions of the most complex task: patient-education (nurses) and obtaining a consent (doctors).

Test Development

Production of Pilot Assessment Instruments

The development of draft items (questions and tasks) and rating scales (or descriptors), which together made up the assessment instruments, was undertaken between November 1999 and March 2000.

Extensive discussion and consultation between the Sydney project team and the authors of the feasibility study took place before final decisions were made about the assessment model and selection of the preferred combination of test options. A meeting was held in Sydney on 21 January, 2000 between the test designer and members of the project team (Clair Mathews, Cathy Noble, Anna Whelan) to discuss draft test specifications and ideas for test tasks. Development of the pilot tests took place in Melbourne, but involved regular communication by e-mail, fax and phone with SWSAHS project staff during February and March to check the suitability of test instruments and tasks, and to seek reaction from bilingual staff in the field.

The pilot tasks and assessment criteria were based on data gathered during the feasibility study (including a review of the literature and existing relevant tests) and scrutiny of translated transcripts of bilingual health interactions gathered by SWSAHS. Design of the assessment instruments took account of constraints on

the content and administration of the tasks/items, as previously identified in the feasibility study. They included explicit instructions for the conduct of the test and the assessment of candidate performance.

Language Background and Self-Assessment Questionnaire

The questionnaire is in five sections (Appendix 1)
The first three seek background information about the participants:

- Section A - Personal Details
- Section B - Employment Details
- Section C - Educational Qualifications (including details on education in the LOTE)

The fourth and fifth sections of the questionnaire involve self-assessment:

- Section D - LOTE Use (including questions on frequency of LOTE-use and self-assessment of general proficiency)
- Section E - Health Professional LOTE Use

The questionnaire is intended for distribution to bilingual staff at the time of appointment in order to find out which might be willing and appropriate subjects for further assessment. For the purpose of the pilot test administration, the questionnaire was to be completed shortly before the interview, and returned to SWASAHS for analysis and comparison with the test scores.

Oral Test Content and Procedures

The pilot test is about 20-25 minutes in duration. The written task instructions are presented in English, on the assumption that all candidates are proficient in English, and also to permit the same test tasks and booklets to be used across a range of different language groups. The interviewers' instructions, however, are delivered in the LOTE (after the preliminary setting-up and checking of test procedures). The mode of delivery chosen for the pilot test was telephone, and the interviews were tape-recorded. The choice of telephone mode was made in the attempt to achieve maximum flexibility in the delivery of the test across different geographical locations and to minimise disruptions to staff. In the event, there were substantial difficulties in organising test administration. (For details see Section: Selection of Candidates)

The test consists of five test tasks, of varying length, complexity and perceived difficulty. Based on Bilingual Health Communication Model, the first four tasks are brief role plays, beginning with a simple social interaction (giving directions) which involves no specialist knowledge or vocabulary. The tasks are designed to range progressively from relatively simple to more complex, from general social communication to more complex/specialized types of professional interaction. They are varied in preparation and performance time, to reflect the increasing complexity and number of exchanges required in each type of interaction. Two versions of the tasks were developed for the pilot test, including alternative versions of the fourth task for doctors and nurses (See descriptions in the following Summary of Tasks.)

The role plays are tightly structured, with each step in the exchange specified in the written instructions for both interviewers and candidates, in the attempt to standardise the content of each task as far as possible. All relevant technical information about procedures is included in the task instructions, in the interests of ensuring that the tasks do not appear to assess candidates' technical knowledge.

The test begins with a warm-up phase of one to two minutes. The purpose of this introductory phase is to establish the candidate's professional status in the test and to put him/her at ease. The interviewer asks a few questions about the candidate's current work – position, department and main areas of responsibility. These are genuinely open-ended enquiries rather than closed (yes/no) questions. This phase is not assessed.

Summary of Tasks

Task 1: Giving directions

The emphasis of this task in both versions is on the candidate's ability to give simple direction and to interact appropriately on a social level with a patient/ client. The assessment of task performance focusses on how well the candidate can handle a simple social exchange and provide clearly comprehensible directions.

Task 2: Explaining procedures

The emphasis of this task is on the candidate's ability to give clear instructions to a patient about a relatively straightforward technical (pre-operative/investigative) procedure. The steps and the respective roles in the interaction are outlined in the task sheets.

The interviewer's input is minimal. In this task, the interviewer is required to interrupt the candidate's explanation so that the candidate is put under some pressure to explain as clearly as possible.

Task 3: History-taking

The emphasis is on the candidate's ability to ask questions and to elicit the appropriate information from a patient while taking a part of the patient's history (in both versions, severe pain). This is an established professional routine with conventional stages which are reflected in the sequence of questions to be asked by the candidate. The interviewer is instructed not to volunteer information, but to respond appropriately to the candidate's questions, using the information supplied on the task sheets.

Task 4: Patient Education (nursing and allied health staff only)

The aim of this task is to see whether the candidate can manage to conduct an extended interaction, which is a complex aspect of their professional work, using the LOTE. The interviewer plays the role of a patient who needs complex instructions about management of a medical condition. The patient's comprehension of the explanation needs to be checked. It is more open-ended than the previous task and therefore a little more unpredictable. All the necessary information is included on the task sheet.

Task 4: Informed consent (medical staff only)

The aim of this task is to see whether the candidate can manage to conduct an extended interaction, which is a complex aspect of their professional work, using the LOTE. The interviewer plays the role of a patient about to undergo a surgical or investigative procedure which entails significant risk, and therefore, legal obligations on the part of the hospital. The patient's informed consent must be obtained after the procedure has been explained by the doctor.

Task 5: Health terminology

This is an experimental task, which has been included in the pilot test in order to assess the candidates' knowledge of health-related vocabulary. It also requires candidates to render the terms, where appropriate, into language which would be understood by a lay person (patient) with no technical knowledge.

In order to see whether or not this task added any significant information about candidates' proficiency not already apparent over the four role-plays, it was decided to pilot the task (of 30 discrete items) and to compare the results with scores on the role play assessments). As pointed out in the feasibility study, the crucial aspect of validation of a test of vocabulary knowledge (as with sample work tasks) is whether or not it is representative of the domain, and this was expected to be unlikely with so small a sample of items.

Pilot Assessment Criteria

The first four test tasks were each assessed separately on a four-point scale (from Level 4, Advanced Professional Competence, to level 1, Ungraded. See Appendix 2) according to two criteria: linguistic and task fulfilment. Using two categories of criteria was thought necessary in order to help assessors distinguish between those features of performance related to language ability and those relating to familiarity with test task and work roles. The final assessment however was to be a holistic or global score, also out of 4, using the band score descriptors for guidance and interpretation. (See Appendix 3 for assessment sheet details.)

Summary of draft band levels and assessment criteria

Level 4 Advanced professional proficiency

At this level, a candidate would be expected to cope with all medical interactions, including informed consent or complex patient education.

Level 3 Professional proficiency

At this level a candidate would be expected to cope with simple or routine medical interactions (such as pre-operative procedures) but not those involving specialised terminology (such as informed consent or complex patient education).

Level 2 Social proficiency
At this level a candidate would be expected to cope with social interactions with patients (such as routine daily conversation or explaining hospital facilities), but not those involving the explanation of medical procedures.

Level 1 Ungraded
At this level, a candidate would not be expected to cope with interaction with patients in the health care context.

Trialling the Pilot Assessment Instruments

Pilot testing and assessment occurred between April and August 2000. This section of the report includes an account of the recruitment of test candidates and the practicalities of test administration, (including recommendations for future improvements). It also covers the selection and training of interviewers and the results of the pilot assessments, including an account of the preliminary validation work to establish standards of test performance.

Range of Pilot Candidates

Candidates were recruited from a range of work contexts, including both health professional and allied health care staff. In addition, a number of support staff (not directly involved in patient care) were included, in order to assess whether or not the test tasks were appropriately targeting proficiency to perform work roles in the LOTE. That is, it was expected that staff who were proficient in the language but not familiar with the work tasks would have difficulty performing them satisfactorily.

Selection of Pilot Test Participants

A list of bilingual health staff was available from the staff surveys of SESAHS conducted in August to December 1999 (SESAHS, 2000). Bilingual staff were asked whether they were willing to let other people at work know about their ability to speak a LOTE. Most staff, 64.7%, indicated that they were willing to let others know. Of those who were not willing to disclose their language skills, the majority indicated their reason as not being able to speak the language well enough (63.8%). Issues relating to job role were also, to a lesser extent, reported: 10.2% responded that it wasn't their job and 9.2% responded that it means extra work for them. Staff were also asked to indicate their interest according to three levels of disclosure of their community language skills:

- only in the department within which they work, 28%;
- being listed on a facility/organisation-wide register, 23.5%;
- being listed on this register and having their language skills reviewed, 38.9%.

A total of 326 staff (38.9%) indicated a willingness to have their language skills reviewed. There were 165 staff who spoke Cantonese but only 11 who spoke Vietnamese. Staff who were interested in being contacted further wrote in their telephone number. Of the 165 Cantonese speaking staff, 117 provided their contact information. This information was cross-tabulated with the staff category.

A list of bilingual staff was available from the staff survey of SWSAHS conducted in late 1996 to early 1997. The main issue for the SWSAHS data was that details from staff including their name and telephone numbers were several years old.

Procedure to Recruit Participants

A list of staff in the selected languages with the varying levels of disclosure was generated. Those who included telephone numbers and consented to being contacted were telephoned at that number by research staff from SWSAHS research group. The target number to recruit for the pilot test was 25 staff in each language. Research staff made several attempts to each telephone contact but were only able to recruit 20 in each language. Most staff who were able to be contacted consented, with only a few declining. The telephone calls were made during the day, which resulted in staff who were on other shifts being missed. Vietnamese-speaking staff were extremely difficult to locate and required the support of multicultural health managers to support the project.

The project was explained and an invitation to assist in developing the pilot test was extended. Most bilingual staff who were contacted consented. They were then asked whether there was a particular time or day which would suit them best to have the test at work. It was explained that the telephone assessment would take approximately 30-40 minutes (including the time needed to complete the questionnaire) and that this time needed to be agreed to by their managers. Staff were asked the names and contact numbers of their managers so that the procedure could be explained to them. Information about the study was sent to the managers by the SESAHS research team who were then to send it through their mailing system. However, numerous problems were encountered in this phase, with some names of managers and units being incorrect. Once these details were clarified, all managers were sent a letter explaining the project, the procedure for testing and the test material for the particular staff member in their unit. It was explained that the staff member needed a quiet place to take the telephone call, some minutes to complete the self-assessment before the telephone call, and time to read the test material, which was to be handed out by the manager (or a delegate). After the test, the material was to be collected and posted back to the research team by the manager.

Cantonese-speaking staff were located mostly from the SESAHS list, as there were larger numbers of such staff within this area and area staff had access to recent contact telephone numbers. Most attempted telephone contacts with staff on the SWSAHS list failed, as bilingual staff involved in the 1996/7 survey had moved on and so were no longer known in the units. There was little success in contacting Vietnamese staff in SESAHS, as so few were listed; efforts to locate Vietnamese-speaking staff in SWSAHS had to be made through multicultural health staff networks.

Once managers had been contacted and agreed to staff involvement in the test, the interlocutors were asked to contact the bilingual participants on the same telephone number and to set up a mutually convenient time to conduct the assessment. At this stage, some misunderstandings occurred, as several of the

interlocutors had assumed that this first contact would be the actual test administration. A number of bilingual staff who had originally agreed to sit the test later proved difficult for the interlocutors to contact. In one case, seven telephone calls were made by the interlocutor before an interview time could be arranged. It is evident that the nature of rostering in the health care system is a difficulty which future test organisation must take into account. Given the voluntary nature of participation in the pilot test, great flexibility was required on the part of the interlocutors.

Recommendations for Future Test Administration

Coordination of test administration

Future tests could be better organised by enlisting the help of multicultural health managers and employee services in the various sites. Once managers have identified suitable bilingual staff, they should seek their consent to be assessed and this information sent on to multicultural health units (or employee services). When a sufficient number have been reached, the managers could contact interlocutors for a specified day or time to conduct the assessments. Alternatively, arrangements could be made to have the telephone tests at work, but at times specified by the managers in consultation with the bilingual staff member. This would avoid the frustration of interlocutors being unable to locate staff and needing to make multiple efforts to contact staff. Another possibility which would avoid the technical problems of recording the telephone interview, would be to conduct as a face-to-face interview at a central test venue. These changes to test administration, or a range of possible combinations, will need further discussion and evaluation of their practicality.

Interviewers

Two interviewer-assessors per language were identified and recruited by SWSAHS staff, with the help of staff from the University of Western Sydney Department of Languages and Linguistics. All four are experienced heath care interpreters and had undergone training to assess language proficiency of volunteers for the Sydney Olympic Games Organising Committee Volunteer Project.

Interviewer Training Workshop

A three-hour training workshop for the four Cantonese- and Vietnamese-speaking interviewers was held on 22nd March, 2000 at the Centre for Allied Nursing Research, Liverpool Hospital. It was conducted by the LTRC test designer, Elisabeth Grove, and the SWSAHS coordinator, Anna Whelan. The training session provided an introduction to the test, to the content of the tasks, use of the draft assessment criteria and test administration procedures. Copies of the test booklets and detailed training notes were also provided for reference. As a result of discussion during the workshop, a number of revisions were made to the pilot tasks, the formatting of the booklets and the draft assessment procedures.

Results of Pilot Tests

It was intended that the test instruments (including the associated scoring/rating procedures) would be piloted on 20-25 health care workers from each of the three categories in five different hospital and community health settings within the South Western and South Eastern Area Health Services between March and June 2000. In the event, the process was more protracted than originally intended, and the total number of pilot candidates smaller than planned (17 for both language groups).

Candidate profile

As previously indicated, for the purpose of the pilot test, candidates were recruited from a range of work contexts and at a range of proficiency levels. The pilot sample also included a number of support staff in order to assess whether or not the test tasks were appropriately targeting proficiency to perform work roles in the LOTE, that is, to help establish the *content validity* of the test. It was expected that staff who were proficient in the language but not familiar with the work tasks would have difficulty performing them satisfactorily.

Candidate profile by language

Cantonese group

As shown in the following table, the majority of test candidates were nurses. There were four allied health staff from a variety of contexts, and four administrative or support staff (one of whom was medical-legal manager with a nursing background). Only one doctor sat the test.

Nurses	8
Medical practitioner	1
Allied Health (unspecified)	1
Pharmacist	1
Physiotherapist	1
Dietary aide	1
Administrative, clerical staff and support staff	4

Because of the incomplete return of questionnaires by test participants (12 out of a total of 17), the data on candidates' backgrounds are limited. (This incompleteness of data has implications for what can be reported about the potential validity of self-assessment.) However, the information available suggests that the participants possessed an appropriate range of experience in Cantonese to make this a suitable sample for the pilot test:

- the majority (8) were born in Hong Kong, two in China, one in Taiwan and one in Australia
- ten of the 12 reported that Cantonese was the language most spoken at home
- the majority (8) had received all or most of their secondary education outside Australia: of these, 6 had been educated bilingually in English and Cantonese (in Hong Kong).
- Five of the seven nurses who had gained their initial professional qualifications in English-medium institutions had studied in Hong Kong

- three had gained their initial professional qualifications in Cantonese
- four had received all of their education in Australia.

Vietnamese group

The majority of candidates tested (7) were nursing trained. As in the Cantonese group, only one of the candidates was a medical practitioner. The main difference from the Cantonese group was the number of multicultural health educators and health workers (some with Vietnamese nursing qualifications).

Nurses	7
Medical practitioner	1
Psychologist	1
Multicultural health workers	2
Multicultural Health Education (VN nursing quals)	1
Multicultural Obstetric Liaison (VN nursing quals)	1
Research officer	1
Administrative, clerical staff and support staff	2

For this group of 17 test candidates, background data on language and educational background is limited. A total of ten questionnaires was returned, but only eight (less than 50% of the whole group) were completed by candidates who actually sat the test. (The extra two respondents had previously volunteered to participate but were unavailable for interview during the period of piloting.)

On the basis of so few responses, the effectiveness of the pilot test can be evaluated only very tentatively. Whether or not the entire group of 17 Vietnamese-speaking participants possessed an adequate range of experience in the LOTE for the purposes of piloting is difficult to ascertain. However, the eight Vietnamese respondents to the questionnaire reported less diverse educational experience in the LOTE than the group of 12 Cantonese, as follows:

- all respondents were born in Vietnam
- seven of the eight reported that Vietnamese was the language most spoken at home
- all had received some or all of their secondary education in Vietnam (between 2 and 13 years)
- all but two had gained their initial professional qualifications in Vietnam.

The incompleteness of the questionnaire data also has implications for the investigating potential validity of candidates' self-assessment of proficiency: with so few cases, there is an inadequate basis for more than preliminary analysis. It may be that the apparently restricted range of participant background in the respondent group was influenced by the fact that several candidates were ethnic health liaison or education officers (not a group for whom the test is ultimately intended, as they are already employed to use their LOTE skills to communicate with patients). However, without more extensive data collection, we cannot be sure.

As will be argued in the next section of the report (Test Results), the existence of a relatively similar range of language background experience in the Vietnamese

group appears to be borne out by the narrower spread of scores awarded by the assessors.

Test Results

In this section, we comment briefly on the range of scores awarded to the candidates by the inteviewer-assessors, and then focus in more detail on a comparison of the assessments by the language-trained and health professional assessors.

Interviewer assessments

A total of 34 candidates (17 Vietnamese- and 17 Cantonese-speakers) were interviewed, each interviewer doing half of the interviews. All but two of the interviews were conducted by telephone. The face-to-face delivery of two Vietnamese interviews occurred because of technical problems with the telephone pick-up recording device. All interviews were audio-taped to allow for:

a) double rating by the interviewers (for reliability of measurement);
b) later assessment by health professional informants (For comments on preliminary validation and standard-setting, see following section).

Assessment procedure

The tapes were first assessed by the interviewer, then handed over to the second assessor for independent assessment. The purpose of the second assessment was to establish whether or not the assessors were interpreting the rating scale in the same way, and if their judgements were consistent (i.e. whether there was reliability between raters). This was intended as a necessary preliminary stage in setting appropriate standards of performance and establishing assessment procedures.

The results of the two assessments were then compared and ten tapes for each language selected for further assessment by two health professional informants (a medical practitioner and a nurse). The role of the specialist informants was to assist in a) setting appropriate performance standards, b) refining the assessment criteria, and c) providing feedback on test tasks, which could be used to further revise the test.

The first four test tasks were each assessed separately on a four-point scale (from Level 4, Advanced Professional Competence, to Level 1, Ungraded. See Appendix 2) according to two criteria: linguistic and task fulfilment. As previously explained, using the two categories of criteria was thought necessary to help assessors distinguish between those features of performance related to language ability and those relating to familiarity with test task and work roles. The final assessment, however, was to be a holistic or global score, also out of 4, using the Level descriptors for guidance and interpretation. Task 5, the test of health terminology, was to be awarded a final score out of 30 for the number of items correct, but this was not to be included in the assessment of the overall score

The scores for each candidate were awarded independently by the assessors, and therefore, exhibit predictable variation. It should be noted here that the draft criteria developed for the pilot test had no external validity – they were based on desiderata arising from discussions with the project team, perusal of reports on their studies, and on the experience of other similar tests. As previously indicated, they were designed in line with the implied hierarchy of the Bilingual Health Communication model:

- to help attune the interviewers to relevant aspects of test performance
- to separate candidates into a small number of relevant levels of performance.

Summary of results

Cantonese group

There was significant agreement between the two Cantonese interviewers. Over the 17 candidates, there was only minor variation between the two raters:

- Overall assessments were identical for 10 of the 17 candidates, and varied by one score point (band level) in 7 cases. (Some of these differences were resolved in the later workshop discussions)
- On individual tasks, there was more variation between raters' scores, but never by more than one score point or band level.
- On Task 5, the test of health terminology, total scores for numbers of correct items were extremely close, varying by no more than 3 points (out of 30). This was an acceptable difference in view of the fact that this section of the test was experimental and that no marking guide had been provided.
- At the pilot stage, this high level of convergent judgement was encouraging, indicating that the Cantonese assessors were interpreting the criteria in a similar way and were generally able to score the tasks without difficulty (This was also clear from discussion in the standard-setting workshop).

Vietnamese group

Technical problems with recording equipment experienced by one of the interviewers resulted in six of the 17 tapes being inaudible; only 11 were available for assessment by the second marker. As a result, there is insufficient evidence on which to base estimates of rater reliability or to comment persuasively on patterns of candidate performance. However, while there was substantial disagreement between the two raters of the Vietnamese group, a consistent pattern began to emerge: one of the raters was more severe than the other in all overall assessments of candidates, and on both linguistic and task fulfilment criteria for each task:

- Differences in overall score were no higher than one band level for any candidate, but the harsher rater awarded an overall Band 4 score to only two of the 11 candidates he assessed
- The more lenient used the Band 4 level for overall test for 13 of the 17 assessments he conducted.

- Of 17 Vietnamese candidates tested, all but two received overall scores of either 3 or 4.

It appears that the candidate sample was not sufficiently spread over a range of proficiency levels to adequately test the effectiveness of the test tasks. (This may also be a product of the high proportion of ethnic health workers included in the pilot sample). It also appears that, at this stage of test development, the judgments of the two raters are too divergent to be used as a reliable guide to candidate performance on the test. Without a larger number of samples and further rater training, it is too early to judge.

Possible reasons for this pattern of discrepancy between the two Vietnamese raters will be discussed in more detail in discussion of the standard-setting workshop. For the moment, it is sufficient to note that the collection of additional samples, a further training session and practice in rating additional samples of performance will be necessary if more satisfactory levels of agreement are to be reached between the Vietnamese assessors. In the event that differences in judgment cannot be resolved, it may be necessary to consider recruiting and training additional raters.

Criteria

In general, the use of the two levels of criteria appears to have been justified by the distinctions made by the interviewer-assessors between linguistic ability and task fulfilment. It was possible for the assessors to award a higher score, where appropriate, to a proficient native speaker, while awarding a lower score for task fulfilment ,where the candidate lacked the professional experience of the task.

Task difficulty

An interesting pattern of task difficulty emerged in the variability of performance across tasks. In summary, somewhat contrary to the intended hierarchy of task difficulty, according to which social interaction and procedures such as giving direction are deemed to be easier than conveying more complex procedural or technical information, a number of candidates had difficulty with Tasks 1 and 2, but less or none with the apparently more difficult tasks 3 and 4. This tends to bear out the perception expressed in the Feasibility Study that the simple/complex dichotomy (simple=social, complex=technical-procedural language) proposed by the Bilingual Health Communication Model might be open to question. More extensive trialling of the revised test will be necessary before the validity of the construct can be commented on with confidence.

Test of health terminology (Task 5)

There was a generally high level of correspondence between the scores awarded to candidates for overall test performance (based on the four role play tasks) and the scores on the final task, the test of health terminology. There was also negligible disagreement between assessors in the scores assigned to Task 5. With one exception, all candidates who scored between 25 and 30 (maximum score) on Task 5 were awarded average overall scores of 3.5 to 4 on the test. At the lower level of proficiency (Level 2, Social competence), two candidates (both born in Australia) scored a total of 8 and 9 out of 30. This would appear to suggest that the vocabulary test may be a good predictor of communicative

ability in the health context. However, another two candidates who received overall scores of 2 to 2.5 also gained very high scores of 27 and 28 respectively on the vocabulary test. One of these candidates was a Taiwanese technical services engineer, who was certainly not involved in clinical work, but working in a hospital setting, may have developed a particular interest in health issues.

With such small numbers of candidates, and limited background information, we cannot be sure of the facts, but it is our impression that the vocabulary test is not a fine-grained enough instrument (nor adequately based on appropriate sampling of the vast field of health-related vocabulary), to have more than the appearance of validity. Nor does it appear to add significantly to the information about candidates which is elicited by their performances on the role plays. It was however, regarded quite favorably by the interviewer-assessors and the health professional informants as an appropriate task which should be retained. Only a couple of minor changes to test items were suggested. For the time being, then, we suggest that:

- Task 5 be retained in the oral test;
- the score for the task not counted formally in determining the score for overall performance); and that
- the performance of candidates on this task be monitored in future test administrations.

Standard-Setting Workshop

Without samples of candidate performance, descriptions of different proficiency levels or standards of performance do not exist except as abstractions The purpose of the combined interviewer-health professional workshop was to define these levels as far as possible, and to begin the work of establishing standards of performance.

Recruitment of health professional informants

For each language, two experienced bilingual health professional informants, a doctor and a nurse, were recruited. All had expressed a particular interest in the issues surrounding the use of bilingual staff within the health service. In one case, the health professional also had post-graduate qualifications in Linguistics as well as experience of clinical teaching.

In preparation for the workshop, the health professionals were asked to assess 10 tapes selected from the pilot group. They were provided with background notes on the test, copies of the interviewers' test booklets and instructions on the use of criteria and scoring procedures.

Assessment procedure

For this exercise, the original assessment sheet was modified, as it had been assumed that the health professionals would find the linguistic/task fulfilment distinction unduly unwieldy. They were asked instead to provided an assessment of performance on each task, using the description of Band levels 1 to 4 as a. An overall assessment was to be provided for each candidate. Task 5 was also to be assessed and the score recorded, but not formally included in the overall

assessment of performance on the test. The occupational experts were also encouraged to comment briefly on any noteworthy features of candidate performance, or to summarise the reasons for their judgements.

Their assessment sheets were returned in advance of the workshop to enable LTRC staff to undertake some preliminary analysis and comparisons between scores.

Results of assessments

In the time available for the workshops (just under three hours), it was possible to listen to and discuss in detail only six of the ten tapes for both the Cantonese and Vietnamese groups. While discussion of a larger number of samples was desirable, it would have been counter-productive to truncate discussion in the interests of greater coverage. Where there were discrepancies of opinion, the reasons for those differences needed to be aired, in the interest of reaching a better understanding among the individual judges

Cantonese group

As previously mentioned, there was a high level of agreement between the two Cantonese interviewers in their assessments of the whole group of 17 pilot candidates. In contrast, on the group of 10 sample tapes chosen for the standard-setting workshop, there was substantial disagreement between the interviewer assessors and the health professional informants. In general, the health professionals tended to be harsher than the interviewer assessors – in most cases, always one band level below. One of the health professional informants was consistently harsher than the other assessors. On all but one of the tapes this assessor was at least one band lower than the others. When it became apparent before the workshop that this assessor was much harsher, she was asked to assess two more tapes. The same pattern of relative harshness was repeated.

This difference between the language-trained assessors and the occupational informants suggests that *the two groups of assessors held different views of the purpose of the test and the nature of the criteria.* In the course of discussion, it emerged that both occupational experts appeared to see their task as essentially involving assessment of the candidates' professional skills in their performance of the tasks. From this perspective, there was no meaningful distinction between the candidates' linguistic and professional communication skills. For the health professional whose judgment was consistently harsher, there was no room for doubt – if a candidate made any error of factor or omitted one detail specified in the task, however linguistically proficient and accurate in performing the rest of the task, the health professional was unwilling to award a score of 4 (advanced professional competence). This decision was justified on the basis of the risk to patients if a candidate made such an error in the real work context. The other informant had similar concerns about the medico-legal implications of awarding the top score to anything less than perfect task performance – it became evident that this informant had understood the task to be an interpreting or translating task, rather than a means of eliciting a sample of the candidates' linguistic skills in order to predict what they might be able to do in the real work situation where they were performing their own duties, not interpreting for someone else or translating written instructions. For this assessor, too, the candidate's

performance on the test task was understood to have direct equivalence to real-world tasks.

The interviewers, on the other hand, who were also experienced health interpreters, were more tolerant of what they perceived as minor errors, and/or as difficulties experienced by candidates as a result of the format or phrasing of a task. In particular, regarding Task 4, the longest and most complex (Patient-Education or Informed Consent), the interviewers felt that there was insufficient preparation time for candidates to read and digest the quite lengthy instructions; they therefore recommended revisions to reduce the information load. Because they had conducted the interviews and experienced the test situation themselves, they were more prepared in a number of cases to attribute less than perfect performance to nervousness or the adverse effects of test conditions, and to recommend changes to those conditions, where necessary, in the interests of maximising candidates' chances to perform to the best of their ability.

Vietnamese group

As previously indicated, one of the Vietnamese interviews was consistently harsher in his judgements of candidates' task performance. He was also more stringent overall than either of the health professional informants. In contrast to the Cantonese group, the main differences in judgement occurred between the two interviewer-assessors. There was a much higher level of agreement between the two occupational experts and one of the raters than for the Cantonese group (on 8 out of the 10 assessments).

In the course of discussion during the workshop, it was mostly possible for the two health professionals to reach consensus with one of the interviewer-assessors, and to make allowances for minor defects in task performance, which were seen to be a product of difficulties with test conditions or task instructions. The harsher rater on the other hand, expressed definite views about what constitutes effective health communication, and was critical of such aspects of candidate performance as lack of enthusiasm or confidence, or inappropriate manner (e.g. too abrupt). In this, his approach to assessment resembled that of the harsher of the Cantonese health professionals.

Comments

1. In view of these fairly persistent differences, which were generally not resolved in the course of the workshop, it is clear that further work will be necessary to set appropriate standards and to refine the assessment criteria. (See final section, Issues and Recommendation.)

2. The time allowed for the standard-setting workshop (just under three hours) was not long enough to allow enough of the tapes to be discussed or persist in attempts to resolve differences of opinion.

3. Given the range of personalities and different backgrounds of the participants, consensus may be impossible to achieve. But with so small a group of assessors, and only ten samples of taped performance reviewed, it would be premature to generalise about the test validity of the plot test or about the reliability of the assessors.

4. There are also sensitive cultural issues (particularly the importance of status and 'face') which need to be taken into account in any standard-setting exercise involving face-to face encounters between member of the same ethnic community, who may also be known to one another in that community context.

However, it is also evident that considerable progress has been made. To have uncovered some of the difficulties in testing and the complexities of bringing together language and occupational experts to judge candidate performance underlines the importance of more precise definition of the purpose of the test, refining the assessment criteria to match that purpose, involving a larger number of *bilingual* occupational experts in discussion of appropriate standards of performance

Self-Assessment

As previously indicated, the small sample size and the incompleteness of the questionnaire data do not provide an adequate basis for evaluating the reliability of self-assessment as a measure of language proficiency (even for the less serious purpose of 'social' language use). Nor is it appropriate to apply statistical methods of analysis to comparisons of candidates' self-assessments and the scores awarded on the test for the small number of candidates (10 per language) whose taped performances were assessed by both the interviewers and the health professional informants. However, where language background and self-assessment data are available for comparison with raters' scores for given candidates, tentative comparisons and comments will be made. Reference is made only to the Cantonese group because of the inadequacy of the limited information available about the Vietnamese participants.

Results

Of the total of 12 questionnaires returned, nine were by candidates whose test performances were assessed by all four assessors. In view of the disagreement between the raters and the specialist informants in assigning scores to candidates, the following comparisons are offered with some caution. However, in general, there was a high level of agreement between the average scores for overall performance on test awarded by the assessors and the candidates' self-assessments of proficiency. It should be remarked, however, that all but two of the respondents were nurses, of whom seven had received their primary nursing training outside Australia, and who would therefore be expected to demonstrate a high level of proficiency in the use of the LOTE for professional communication.

Seven of the eight respondents to the questionnaires self-assessed at the highest level of general proficiency ('I can talk about anything I want to, including aspects of my work'), while only one (the candidate born in Australia and educated entirely in English) rated herself at the next level of general proficiency ('I can talk about most things, including some aspects of my work') All of those who self-assessed highly in terms of general communication skills also tended to rate themselves highly on Section E of the questionnaire, 'Professional LOTE use'. Reponses to the questions which they considered appropriate to their work roles were either at the highest level (Very well) or at the second level (fairly

well) (tending to select the option N/A for those which were not relevant for their work roles) for all but one of the overseas trained nurses.

Recommendations and Issues for Consideration

1. Test validation

It is important to keep in mind that this is a pilot test, which is still in the preliminary stages of validation. Validation of a test is a complex process, involving a range of activities (in the early stages, amassing sufficient data on both candidates and assessors in order to determine that it is testing what it purports to). The information that the first administration of the test provides is therefore a preliminary but essential guide to the appropriateness of the test tasks and the practicality of test procedures. The indications are that the test is both appropriate and practicable. It will, however, be necessary to continue accumulating relevant information on the performance of candidates and assessors in order to assess the effectiveness of the test and the appropriateness of test materials. The test validation process was also constrained by the small numbers available for testing in Cantonese and Vietnamese, and this will undoubtedly also be the case for other relevant languages.

Recommendations

a) Additional candidates should be sought and the data from their performances included in the dataset for further analysis to establish the effectiveness of the test for both Vietnamese- and Cantonese-speaking staff.

b) Funding should be sought for follow-up investigations, for instance, via individual case studies, interviews with test candidates and patients, to assess the predictive validity of test scores, exploring the relationship between candidates' test performance and actual use of LOTE skills in the professional health encounter.

2. Test standards

Divergent views of test purpose and the resulting discrepancies among the interviewer-assessors and occupational experts were to be expected in the pilotting stage. However, it will be necessary, to reconvene both groups and/or to seek the participation of additional experts, linguistic and occupational, in the effort to establish the proficiency levels which the test seeks to measure

It should be noted here that *standard-setting and validation of the test need to be undertaken separately for each language group* - standards of performance are not inherent in the test tasks themselves, but in the behaviour and judgements of those who conduct the tests and perform the assessments. While it is hoped eventually to establish *common standards of performance* across different language groups, extensive investigation and comparative studies will be necessary.

Once standards of performance on the test have been established in each language, further training and practice in assessing additional samples of performance will be necessary if more satisfactory levels of agreement are to be reached, especially between the Vietnamese assessors. In the event that differences in judgment cannot be resolved, it may be necessary to consider recruiting and training additional raters.

Recommendations

a) Further meetings should be convened between interviewers and occupational experts in Cantonese and Vietnamese to establish proficiency levels on the test in both languages.

b) Comparative investigations of performance on the test across different language groups should be undertaken as part of the extension of the test to other languages.

c) LTRC involvement in the extension of the test to other languages should be maintained, in the interests of maintaining test consistency across languages and permitting the establishment of a comprehensive data set which may be used in the future to investigate the validity of the test across languages.

d) Transcription of a sample of interview tapes should be undertaken to establish consistency of interviewer behaviour within and across languages.

e) Re-rating of taped speaking test interviews by independent LOTE experts would assist in establishing the reliability of the interviewer assessments.

3. Test candidates

The numbers of candidates in both language groups were small, and the range of backgrounds relatively restricted (for Vietnamese in particular). This may be a reflection of the small numbers of health professionals from these groups currently employed in the Area Health Services in question. It may also be a sign of some reluctance among staff to undertake the test. The low rate of return of the questionnaires could also indicate concern among relevant staff about how the information they provide will be used by the employer. It will therefore be most important to ensure that staff are reassured on these matters and that the test purpose is explained when they are approached to sit the test. How test results are to be reported to both candidates and Area Health Service is a still to be determined and cannot be finalised until the test has been adequately trialled and evaluated.

Recommendations

a) Further consideration should be given to the suitability of the content of the test questionnaire, via feedback from bilingual staff.

b) Efforts should be made in future tests to ensure that staff who sit the test have previously completed the questionnaire and that it is returned with the test results for comparative analysis of data.

4. Test materials and procedures

Feedback from the interviewers and health professional informants has proved useful in informing revisions to test tasks and formatting, especially Task 4 of the oral test. For future test administrations, both in the languages already tested and those to come, revised versions will be used and their effectiveness evaluated. Continuity of personnel and continued contact with the LTRC test developers will be important in establishing common procedures and standards.

Recommendations

a) For the extension of the test to other languages, the same (revised) version of the test should be used as for Cantonese and Vietnamese.

b) Cantonese and Vietnamese interviewers should be invited to participate in training sessions for interviewers in other languages, in order to share experiences and expertise gained in the first test administration.

Recommendations for Future Test Administration

- As previously indicated, adjustments need to be made, in particular, to the arrangements for test administration and recruitment of candidates. Test administration could be streamlined by enlisting the help of multicultural health managers and employee services in the various sites.

- The use of the taped telephone interview as the mode of test delivery needs further investigation. In the extension of the test to other languages, consideration should be given to the use of both face-to-face and telephone mode in order to maximise flexibility for both interviewers and test takers.

References

Brown, A (1993) LSP testing: the role of linguistic and real-world criteria. *Melbourne Papers in Language Testing 2,2*: 35-54

Davies, A. (1995) Testing communicative language or testing language communicatively: what? How? *Melbourne Papers in Language Testing 4,1*: 1-20.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., McNamara, T (1999) *Dictionary of Language Testing*. Cambridge: Cambridge University Press.

Douglas, D. (2000*) Assessing Languages for Specific Purposes.* Cambridge: Cambridge University Press.

Grove, E. & Brown, A.(1999) *Bilingual Health Workers Language Assessment Project Feasibility Report.* Prepared for the South Western Sydney Area Health Service: Language Testing Research Centre, The University of Melbourne.

Johnson, M., Noble, C., Matthews, C., & Aguilar, N. (1998). Towards culturally competent health care: Language use of bilingual staff. *Australian Health Review*, 21(3), 49-66.

Johnson, M., Noble, C., Matthews, C., & Aguilar, N. (1999). Bilingual communicators within the health care setting. *Qualitative Health Research,* 9(3), 329-343.

Jones, R. L. (1985) Second Language Performance Testing: an overview. In Hauptmann et al, *Second Language Performance Testing*. Ottawa: University of Ottawa Press.

McNamara, T.F. (1990) *Assessing the second language proficiency of health professionals*. Unpublished PhD thesis, University of Melbourne.

Richards, J., Platt, J., & Platt, H. (1993) *Dictionary of Language Teaching and Applied Linguistics*. (3rd Ed.) London: Longman.

Appendix 1:
LANGUAGE BACKGROUND QUESTIONNNAIRE: MEDICAL, NURSING AND ALLIED HEALTH STAFF

PURPOSE

The purpose of this questionnaire is find out details of your background in a Language Other Than English (LOTE), your use of the LOTE and your opinions about your ability to use this language in your communication at work. It follows up the recent survey of staff language skills in which you were recently involved.

Participation in this new survey is voluntary. The information you provide is intended to encourage you to judge your own abilities in the LOTE, and to use them appropriately in your daily communication with patients. If you consider that your language skills are advanced enough for all or most of your necessary communication with LOTE-speaking patients, you may be asked if you wish to have these skills tested more formally.

Please complete using BLOCK LETTERS. Use ticks √ where appropriate in the boxes provided.
Please write **N/A** next to any question which is not applicable to you, and move on to the next.

SECTION A: PERSONAL DETAILS

Family name: …… Given names: ……..
Date of birth: ……. Country of birth: ……
Languages other than English spoken: ……
Date of arrival in Australia (if applicable): ……

SECTION B: EMPLOYMENT DETAILS

Staff category: ……
Name of Hospital or Centre: ……..
Service/ Department:…… Ward (if applicable)……..
Work telephone number: ……
Main area of practice/work? (eg medical, surgical, primary health nursing, mental health etc)
……
Numbers of years of employment as a health/allied health professional? ……

SECTION B: EDUCATIONAL QUALIFICATIONS

We are interested in whether your education was all in The LOTE, all in English, or in a combination of both. Please fill in the details of your main qualifications in the spaces below:

1. Tertiary and professional

(i) Name of qualification: ……
Name and place of institution: ……
Language of instruction: ……
Date of completion: ……

(ii) Name of qualification: ……
Name and place of institution: ……
Language of instruction: ……
Date of completion: ……

(iii) Name of qualification: ……
Name and place of institution: ……
Language of instruction: ……
Date of completion: ……

Any other relevant information

2. Secondary education

1. How many years of your secondary education were in the LOTE? …….. years
2. How many years of your secondary education were in English? …….. years
3. Did you study the LOTE as a school subject? YES / NO (circle)
4. If YES, for how many years? …….. years
5. Any other relevant information

……

3. Primary education

1. How many years of your primary education were in the LOTE? …….. years

2. How many years of your primary education were in English? …….. years
3. Any other relevant information

……

SECTION D: LOTE USE

We are interested to find out how you use your LOTE language skills.

1. Which language do you use more at home? English OR the LOTE (circle)

The following questions concern how often you speak the LOTE in different contexts.

Please tick √ the box next to the most appropriate phrase after each question.

2 How often on average do you speak the LOTE at home?

ρ Every day ρ Several times per week ρ Several times per month ρ Occasionally ρ Never

3. How often on average do you speak the LOTE outside the home (but not at work)?

ρ Every day ρ Several times per week ρ Several times per month ρ Occasionally ρ Never

4. How often on average do you use the LOTE at work to communicate with colleagues?

ρ Every day ρ Several times per week ρ Several times per month ρ Occasionally ρ Never

5. How often on average do you use the LOTE at work to communicate with patients?

ρ Every day ρ Several times per week ρ Several times per month ρ Occasionally ρ Never

6. How would you describe your ability to speak the LOTE?

Please tick only ONE of the boxes next to the following statements.

- ρ I can talk about anything I want to, including aspects of my work
- ρ I can talk about most things, including some aspects of my work
- ρ I can talk on most topics of daily conversation
- ρ I can talk on a few very simple topics of conversation
- ρ I can say a few simple things (greetings, asking the time, commenting on the weather etc)
- ρ I am not confident of my speaking ability, but I can understand more than I can say
- ρ I cannot speak the LOTE at all, but I have a general understanding of the culture and community attitudes.

Any other relevant information……

7. How would you describe your ability to write in the LOTE?

Please tick only ONE of the boxes next to the following statements.

- ρ I can write accurately in the LOTE about anything I want to, including aspects of my work
- ρ I can write fairly accurately in the LOTE about most things, including some aspects of my work
- ρ I can write fairly accurately in the LOTE on a range of general but not professional topics
- ρ I can write in basic LOTE on simple topics (eg short personal letters)
- ρ I can write simple sentences in the LOTE for very limited purposes (eg thank you notes, New Year messages)
- ρ I can write only a few words of the LOTE
- ρ I cannot write in the LOTE at all.

Any other relevant information……

……

……

……

SECTION E: HEALTH PROFESSIONAL LOTE USE

We are interested to know what you think of your own ability to communicate with LOTE-speaking patients. This concerns not only what you do now, but also what you **think you could do** in the LOTE in working with patients.

How **well** do you think you could carry out the following types of communication in the LOTE? Please tick the box next to the most appropriate answer beneath each item in the following list. Choose N/A for any that are not applicable to your work

Conversation on simple general topics (eg greetings, weather)

ρ Very well ρ Fairly well ρ With difficulty ρ Not at all ρ N/A

Giving simple instructions or directions

ρ Very well ρ Fairly well ρ With difficulty ρ Not at all ρ N/A

Finding out how a patient is feeling

ρ Very well ρ Fairly well ρ With difficulty ρ Not at all ρ N/A

Reassuring a distressed patient/client

ρ Very well ρ Fairly well ρ With difficulty ρ Not at all ρ N/A

Taking a case history

ρ Very well ρ Fairly well ρ With difficulty ρ Not at all ρ N/A

Conducting an assessment of a patient's/client's condition

ρ Very well ρ Fairly well ρ With difficulty ρ Not at all ρ N/A

Explaining a diagnosis/prognosis

ρ Very well ρ Fairly well ρ With difficulty ρ Not at all ρ N/A

Identifying problems and giving explanations

ρ Very well ρ Fairly well ρ With difficulty ρ Not at all ρ N/A

Explaining a technical procedure

ρ Very well ρ Fairly well ρ With difficulty ρ Not at all ρ N/A

Explaining treatment options

ρ Very well ρ Fairly well ρ With difficulty ρ Not at all ρ N/A

Checking a patient's understanding of a treatment plan

ρ Very well ρ Fairly well ρ With difficulty ρ Not at all ρ N/A

Providing counselling/patient education

ρ Very well ρ Fairly well ρ With difficulty ρ Not at all ρ N/A

Obtaining informed consent to a procedure

ρ Very well ρ Fairly well ρ With difficulty ρ Not at all ρ N/A

THANK YOU FOR COMPLETING THIS QUESTIONNAIRE

Appendix 2: BILINGUAL HEALTH COMMUNICATION SKILLS TEST ASSESSMENT SCALE

Level 4 Advanced professional proficiency

Candidates at this level have a high proficiency in the language, such that they would be expected to be able to cope effectively and confidently with the full range of tasks, including those involving specialist vocabulary. They are able to converse freely and fluently, and have a broad vocabulary which enables them to use lexical choice to good effect. They are able to use the language accurately and expressively by drawing on a broad knowledge of grammar as well as idioms, colloquialisms and cultural references. Speech is fully intelligible.

At this level, a candidate would be expected to cope with all medical interactions, including informed consent or complex patient education.

Level 3 Professional proficiency

Candidates at this level would have sufficient proficiency in the language to undertake most tasks, although they would still be expected to have some problems with those requiring the use of specialised vocabulary. Speakers at this level can almost always express ideas well and rarely have to grope for words or ask for clarification. Occasional errors do not interfere with communication.

At this level a candidate would be expected to cope with simple or routine medical interactions (such as pre-operative procedures) but not those involving specialised terminology (such as informed consent or complex patient education).

Level 2 Social Proficiency

Candidates at this level would generally have sufficient competence in the language to be able to undertake simple or routine tasks, such as social interaction with patients and dealing with non-medical problems. Their vocabulary is such that while they are able to communicate reasonably well on social topics, they are not able to communicate on those requiring any specialised vocabulary. They are able to converse in a participatory fashion and are generally able to express ideas confidently, although not necessarily with ease. Errors rarely interfere with understanding, although under stress complicated structures break down.

At this level a candidate would be expected to cope with social interactions with patients (such as routine daily conversation or explaining hospital facilities), but not those involving the explanation of medical procedures.

Level 1 Ungraded

Candidates at this level would generally have difficulty using the language to communicate in the health care context. They do not have the range of vocabulary and expression to converse on areas of occupational importance. They have difficulty imparting even simple information, description or instructions. Lack of fluency and comprehension is a barrier to

effective and easy communication; interacting with such a speaker would cause strain for the native speaker.

At this level, a candidate would not be expected to cope with interaction with patients in the health care context.

Appendix 3: BILINGUAL HEALTH COMMUNICATION SKILLS TEST

Candidate_____Interviewer _____

Profession_____Language_____Test version Date_____

Tick relevant space (4, 3, 2 or 1) for both judgements in Tasks 1-4.

Task 1	**4**	**3**	**2**	**1**
Linguistic assessment				
Task fulfillment assessment				
Task 2	**4**	**3**	**2**	**1**
Linguistic assessment				
Task fulfillment assessment				
Task 3	**4**	**3**	**2**	**1**
Linguistic assessment				
Task fulfillment assessment				
Task 4	**4**	**3**	**2**	**1**
Linguistic assessment				
Task fulfillment assessment				

Task 5 Total number of items correct Score _____

Final assessment Level _____

Comments:

……..