OET OCCUPATIONAL ENGLISH TEST

Occupational English Test

THE UNIVERSITY OF MELBOURNE

OCCUPATIONAL ENGLISH TEST

An overview of the Occupational English Test

The purpose of the Occupational English Test (OET) is to evaluate the English language competence of people who have trained as medical and health professionals in one country (e.g., India) and wish to gain provisional registration to practise their profession in an English-speaking context (e.g., Australia). In most cases, applicants are subsequently required by their professional board or council to sit a test of professional knowledge.

Candidates for the OET are from 12 health professions: dentistry, dietetics, medicine, nursing, occupational therapy, optometry, pharmacy, physiotherapy, podiatry, radiography, speech pathology, and veterinary science. Nursing, dentistry and medicine currently provide the largest numbers of candidates.

The test is administered by the OET Centre, a business unit of the Centre for Adult Education (CAE) in Melbourne, Australia, which assumed responsibility for the test in 2004.

All four language skills are tested – listening, reading, writing and speaking – with an emphasis on contextualised communication for professional purposes. The Speaking and Writing sub-tests are specific to each profession, while the Listening and Reading sub-tests are common to all candidates.

The test is currently used by the governing bodies of the professions at state and national level in Australia and New Zealand and by the Australian Department of Immigration and Citizenship (DIAC). Each board or council determines the results required from candidates to meet the language competency standards to function in the profession.

The OET uses a secure test bank from which materials are selected for each administration. The only past test materials available are published by the OET Centre. Listening and reading test materials are developed for the OET Centre by staff at the Language Testing Research Centre (LTRC) of the University of Melbourne. LTRC staff also carry out the analysis of test data. From time to time specialist test consultants are engaged to devise further test materials in line with the same technical processes.

Writing and Speaking materials for the test are written in consultation with professional educators for the regulatory authorities of the individual professions or with experienced practitioners.

OET Statements of Results include a scaled band score (grade) for each of the four sub-tests. Scores are considered valid for two years by most professional regulatory bodies.

The OET Centre does not currently set a limit on the number of times a candidate may present for testing. Many candidates may re-sit the sub-tests for which they do not obtain a satisfactory grade without re-sitting those for which the grade is satisfactory. Some professional regulatory authorities, however, require candidates to obtain satisfactory grades on all four sub-tests at one administration of the test.

Frequency and location

The OET is currently administered seven times a year in over 40 locations around the world. The largest testing centres are in Australia. From 2010 the number of test administrations will increase to ten.

Security

Test materials are sent to test centres by secure courier. Everyone involved in administering the test signs a confidentiality agreement. All test materials are returned to the OET Centre in Melbourne by secure courier and accounted for.

Administrators

The OET is administered at each venue under the direction of a Venue Co-ordinator. Detailed instructions explaining the required administration procedures are included with the test materials and Co-ordinators check that all standards required by the OET Centre are met. Interviewers have native or native-like competence in English.

Assessment

Assessors must have a tertiary qualification or equivalent and a TESOL qualification. A post-graduate qualification in language testing is desirable. Assessors undergo specific training in OET assessment methodology.

Assessment is conducted at the OET Centre in Melbourne and preceded by training and standardisation for assessors to achieve optimum consistency of standards.

Assessment of the Reading sub-test – a multiple-choice questions (MCQ) test - is computerised and is analysed initially by the University of New South Wales and then further by the University of Melbourne for performance of individual items.

The Listening sub-test is assessed against a detailed marking guide prepared by the test designers. Problematic scripts are dealt with as a group by an experienced assessor and all critical borderline scripts are double-marked.

Writing scripts and Speaking interviews are rated twice, with aberrant and unusual cases marked a third time. Assessors use a set of criteria to rate candidates' performance. Analysis of rater consistency and severity is conducted using multi-faceted RASCH analysis.

Registration procedures

All test applications are taken online through the OET website. Candidate photos are uploaded to the website and payment is taken online by credit card. When applying, candidates must also give the exact details of the identity document they will provide when sitting the test. The OET Centre only accepts passports as proof of identity but makes an exception for candidates with a national identity card, provided they are sitting the test in the country that issued it. Candidates undergo a rigorous ID check upon initial registration on the test day and before each sub-test. The ID check procedures have DIAC approval.

Communication with candidates is by email and via the website. Candidates can download a complete set of preparatory materials. Specific details for particular venues and timetables for the individual sub-tests are provided two weeks prior to the day of the test. Candidate information, including photo and identity document number, is sent to the venues for identity verification on the test day.

Special provision

Candidates with special needs are provided for. Specifications for test centres are given in detailed instructions to Venue Co-ordinators. All test centres must comply with local requirements for meeting the needs of people with disabilities.

The OET Centre makes all reasonable arrangements to accommodate special visual or auditory needs, including enlargement of print texts and special auditory equipment.

History of the test

The Occupational English Test was designed by Professor Tim McNamara of the University of Melbourne under contract to the Australian Federal Government.

As part of the annual intake of refugees and immigrants, hundreds of overseas-trained health professionals were entering Australia by the mid to late 1980s. The majority were medical practitioners but a number of other health professional groups were also represented.

The process of registration to practise in most health professions in Australia included three stages of assessment: English language proficiency, a multiple choice test of profession-specific clinical knowledge, and a performance-based test of clinical competence. Dissatisfaction with the results of existing language tests led to the development of thoroughly researched specifications for a communicative, contextualised test. The OET has been frequently reviewed and analysed in the literature over the past 20 years (see references section for examples). McNamara (1996) gives a full account of the development of the test and associated validation research.

The initial development of the test specifications involved:

a. extensive consultation with expert informants, including clinical educators, ESL teachers offering language support in clinical settings, and overseas-trained professionals who were completing or had completed a clinical bridging program.

b. literature search.

c. direct observation of the workplace.

Stages of Test Development, presentation by Prof Tim McNamara, August 2007

STAGES OF TEST DEVELOPMENT

- *Analysing target domain via 'job analysis'*

Job analysis (1): Consultation with expert informants

- **Clinical educators** (especially those with experience of overseas-trained health professionals) [via interview]
- **ESL teachers** offering language support in clinical settings [via interview]
- **Overseas-trained health professionals** who are in or have completed a clinical bridging programme [via questionnaire]

Job analysis (2 and 3):

- *Literature search*
- *Direct observation of the workplace*

Commonalities observed: (1) stages (some optional) of the 'consultation'

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

This event provided the basis for specifications for a common format for assessment of *speaking* and *listening* skills.

Commonalities observed (2)

- Listening:
  - professional development
- Reading:
  - Professional development
  - Case history notes
- Writing:
  - Letters of referral

STAGES OF TEST DEVELOPMENT

- *Identifying resources/constraints: Needed to maximize both*
  - ***Specificity**: plausibility in professional context*

  and
  - ***Commonality**: limits to affordability of profession-specific materials*
- *Drafting of test specifications*

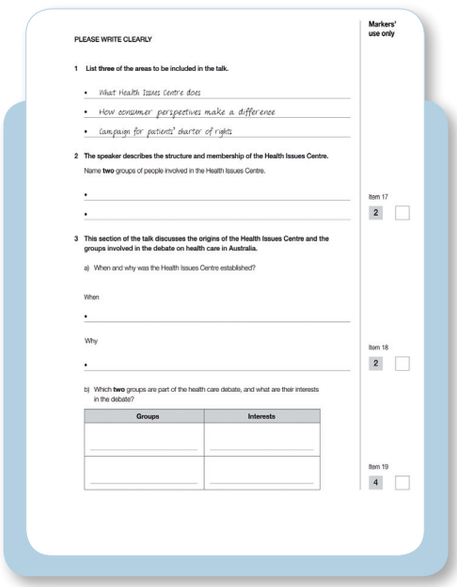
Description of the OET

Test format

The OET assesses listening, reading, writing and speaking.

There is a separate sub-test for each skill area. The Reading and Listening sub-tests are designed to assess the ability to understand written and spoken English in contexts related to general health and medicine. The sub-tests for Reading and Listening are not specific to any single profession but are based on topics and tasks common to all professions.

The Speaking and Writing sub-tests are specific to each profession and are designed to assess the ability to use English appropriately in a relevant professional context.



PLEASE WRITE CLEARLY

Markers' use only

1 List three of the areas to be included in the talk.

- What Health Issues Centre does
- How consumer perspectives make a difference
- Campaign for patients' charter of rights

2 The speaker describes the structure and membership of the Health Issues Centre.

Name **two** groups of people involved in the Health Issues Centre.

Item 17 2

3 This section of the talk discusses the origins of the Health Issues Centre and the groups involved in the debate on health care in Australia.

a) When and why was the Health Issues Centre established?

When

Why

Item 18 2

b) Which **two** groups are part of the health care debate, and what are their interests in the debate?

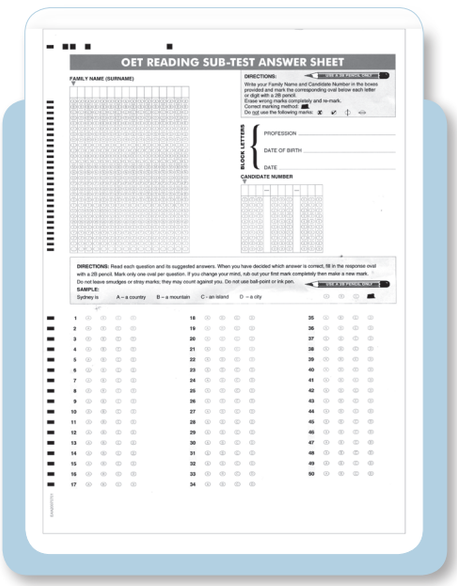
Groups	Interests

Item 19 4

Listening Sub-test

The Listening sub-test consists of two parts: a recorded, simulated professional-patient consultation with note-taking questions (Part A), and a recorded talk or lecture on a health-related topic with short-answer/note-taking questions (Part B), each about 15 minutes of recorded speech. A set of questions is attached to each section and candidates write their answers while listening. The original recording is edited with pauses to allow candidates time to write their answers.

The format for Part A (the consultation) requires candidates to produce case notes under relevant headings and to write as much relevant information as possible. Most questions in Part B (the lecture) include indications on the number of points a candidate is expected to include.



OET READING SUB-TEST ANSWER SHEET

FAMILY NAME (SURNAME)

DIRECTIONS:

PROFESSION

DATE OF BIRTH

DATE

CANDIDATE NUMBER

BLOCK LETTERS

DIRECTIONS: Read each question and its suggested answers. When you have decided which answer is correct, fill in the response oval with a 2B pencil. Mark only one oval per question. If you change your mind, rub out your first mark completely then make a new mark. Do not leave smudges or stray marks, they may count against you. Do not use ball-point or ink pen.

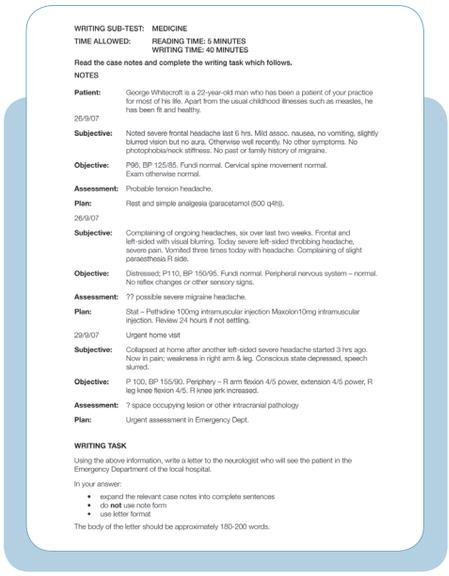
SAMPLE:

Sydney is A – a country B – a mountain C – an island D – a city

Reading Sub-test

The Reading sub-test is made up of two texts dealing with the scientific aspects of a health issue, each followed by a set of multiple-choice questions. The number of questions may vary between 21 and 24.

All questions have been analysed for maximum item reliability and to permit effective discrimination between candidates.



WRITING SUB-TEST: MEDICINE
TIME ALLOWED: READING TIME: 5 MINUTES
WRITING TIME: 40 MINUTES

Read the case notes and complete the writing task which follows.

NOTES

Patient: George Whitecroft is a 22-year-old man who has been a patient of your practice for most of his life. Apart from the usual childhood illnesses such as measles, he has been fit and healthy.

26/9/07

Subjective: Noted severe frontal headache last 6 hrs. Mild assoc. nausea, no vomiting, slightly blurred vision but no aura. Otherwise well recently. No other symptoms. No photophobia/neck stiffness. No past or family history of migraine.

Objective: P96, BP 125/85. Fundi normal. Cervical spine movement normal. Exam otherwise normal.

Assessment: Probable tension headache.

Plan: Rest and simple analgesia (paracetamol (500 q4h)).

26/9/07

Subjective: Complaining of ongoing headaches, six over last two weeks. Frontal and left-sided with visual blurring. Today severe left-sided throbbing headache, severe pain. Vomited three times today with headache. Complaining of slight paraesthesia R side.

Objective: Distressed; P110, BP 150/95. Fundi normal. Peripheral nervous system – normal. No reflex changes or other sensory signs.

Assessment: ?? possible severe migraine headache.

Plan: Stat – Pethidine 100mg intramuscular injection Maxolon10mg intramuscular injection. Review 24 hours if not settling.

29/9/07 Urgent home visit

Subjective: Collapsed at home after another left-sided severe headache started 3 hrs ago. Now in pain; weakness in right arm & leg. Conscious state depressed, speech slurred.

Objective: P 100, BP 155/90. Periphery – R arm flexion 4/5 power, extension 4/5 power, R leg knee flexion 4/5. R knee jerk increased.

Assessment: ? space occupying lesion or other intracranial pathology

Plan: Urgent assessment in Emergency Dept.

WRITING TASK

Using the above information, write a letter to the neurologist who will see the patient in the Emergency Department of the local hospital.

In your answer:

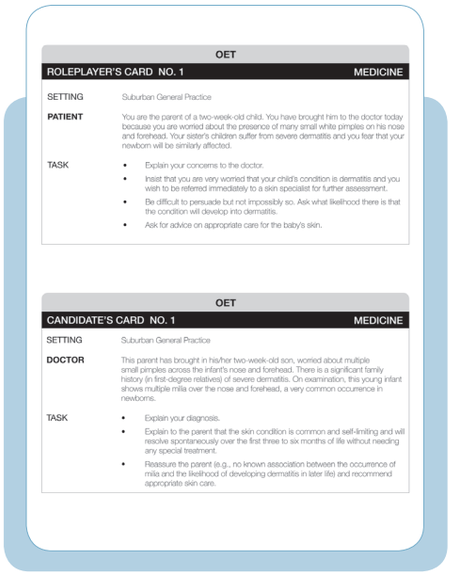
- expand the relevant case notes into complete sentences
- do **not** use note form
- use letter format

The body of the letter should be approximately 180-200 words.

Writing Sub-test

The Writing sub-test usually consists of a scenario presented to the candidate which requires the production of a letter of referral to another professional. The letter must record treatment offered to date and the issues to be addressed by the other professional. The letter must take account of the stimulus material presented.

The body of the letter must consist of approximately 180-200 words and be set out in an appropriate format. For certain professions, other professional writing tasks of equivalent difficulty may also be set, e.g., responding in writing to a complaint, or providing written information to a specified audience in the form of a letter. There is currently discussion with regulatory authorities about expanding the writing task options.



OET

ROLEPLAYER'S CARD NO. 1 MEDICINE

SETTING Suburban General Practice

PATIENT You are the parent of a two-week-old child. You have brought him to the doctor today because you are worried about the presence of many small white pimples on his nose and forehead. Your sister's children suffer from severe dermatitis and you fear that your newborn will be similarly affected.

TASK
- Explain your concerns to the doctor.
- Insist that you are very worried that your child's condition is dermatitis and you wish to be referred immediately to a skin specialist for further assessment.
- Be difficult to persuade but not impossibly so. Ask what likelihood there is that the condition will develop into dermatitis.
- Ask for advice on appropriate care for the baby's skin.

OET

CANDIDATE'S CARD NO. 1 MEDICINE

SETTING Suburban General Practice

DOCTOR This parent has brought in his/her two-week-old son, worried about multiple small pimples across the infant's nose and forehead. There is a significant family history (in first-degree relatives) of severe dermatitis. On examination, this young infant shows multiple milia over the nose and forehead, a very common occurrence in newborns.

TASK
- Explain your diagnosis.
- Explain to the parent that the skin condition is common and self-limiting and will resolve spontaneously over the first three to six months of life without needing any special treatment.
- Reassure the parent (e.g., no known association between the occurrence of milia and the likelihood of developing dermatitis in later life) and recommend appropriate skin care.

Speaking Sub-test

The production of contextualised professional language is achieved by requiring the candidate to engage with an interviewer who plays the role of a patient or a patient's carer. The candidate must respond as a professional consultant to two different scenarios played out with the interviewer. These exchanges are recorded for subsequent assessment. The recording also includes a short 'warm-up' that is part of the interview, though this material is not assessed.

Scoring the test

OET grades are reported on an official report form, the Statement of Results. A band score is reported for each of the sub-tests. These band scores range from A (highest) to E (lowest). They are derived differently for the sub-tests for the productive skills (Writing and Speaking) and receptive skills (Listening and Reading).

Speaking and Writing

The Speaking and Writing sub-tests are marked by trained, experienced assessors based in Melbourne. Each candidate's performance is marked by two assessors, who mark independently of each other and without knowledge of a candidate's performance on the other sub-tests. The Writing and Speaking sub-tests are each graded against five criteria; each criterion has six grade levels, 1-6, with level 6 representing a very strong response.

The criteria for each sub-test are:

Speaking

- Overall Communicative Effectiveness, Intelligibility, Fluency, Appropriateness of Language, Resources of Grammar and Expression

Writing

- Overall Task Fulfilment, Appropriateness of Language, Comprehension of Stimulus, Linguistic Features (Grammar & Cohesion), Presentation Features (Spelling, Punctuation & Layout)

The data are analysed using multi-faceted RASCH analysis (McNamara 1996) with FACETS software (Linacre 1989); candidate and rater are facets in the analysis.

All candidates who are found to have unexpected responses in the data analysis are third marked. To compensate for any differences in assessor severity, band scores are derived from the single fair score generated by FACETS, rather than from averaged raw scores. These fair scores are then converted to bands as follows:

Conversion to band scores (range 1-6)

Band A: 5.6 and above
Band B: 4.8 – 5.5
Band C: 4.2 – 4.7
Band D: 3.4 – 4.1
Band E: 3.3 and below

Reading and Listening

The Reading sub-test is a multiple-choice questions (MCQ) test; score sheets are computer scanned. The item-level data are analysed using a RASCH analysis programme, QUEST, for overall internal consistency and item quality. Any items found to be performing unacceptably (i.e., with fit of above 1.3 or with discrimination levels of less than .25) are removed from the subsequent analysis.

The Listening sub-test is marked by a small group of experienced raters in Melbourne. The test consists of short-answer questions and is marked according to detailed marking guidelines prepared by the test designers. In order to ensure consistency of marking, prior to the marking session at each administration assessors are trained in how to interpret and apply these guidelines. As for the Reading sub-test, the item-level data are analysed using QUEST for overall internal consistency and item quality. Again, any items found to be performing unacceptably (i.e., with fit of above 1.3 or with discrimination levels of less than .25) are removed from the subsequent analysis.

Band cut-scores (i.e., the boundaries between the band scores) are re-set for the Reading and Listening sub-tests at every administration regardless of whether they are new tests, re-constituted tests using two texts/parts which have not previously been used in combination, or previously-used tests.

This is because for every administration, reliability of measurement is maximised by removing from the analysis those items which are found to be performing unacceptably for that cohort. This means that, for the same test version or individual text/part, the number of items may vary from administration to administration, although experience has found that no more than one item is usually removed.

Cut-scores for each band level on the Reading and Listening sub-tests are set on the basis of the percentage distribution of candidates into band levels using the average of the Writing and Speaking sub-tests.

This assumes that normally these criterion-referenced sub-tests should be equivalent in difficulty and that distribution into grades across the sub-tests should therefore be similar for the whole cohort. It does not assume that individuals will get the same grade on each sub-test. (The proportion of candidates falling into in each grade is generally very similar for the Speaking and Writing sub-tests.)

Monitoring assessor performance

Assessors' performance on the Speaking and Writing sub-tests is routinely monitored and reported back to them after every test administration. Assessors are given feedback on two aspects of their performance: the degree to which their ratings are consistent within themselves (their 'fit' estimates in the FACETS analysis) and their severity relative to the other assessors.

Consistency and severity are reported using the following scales:

Consistency

- good
- not enough variation, tendency to rate too conservatively
- inconsistent

Severity

- appropriate
- slightly severe (on average quarter to half a band below mean)
- very severe (on average more than half a band below mean)
- slightly lenient (on average quarter to half a band above mean)
- very lenient (on average more than half a band above mean)

Assessors must perform strongly on these two scales in order to remain in the pool of markers. Further training is offered to assessors when required.

Test procedures

Final safeguard of grades

To guard against error in this complex process, all sets of results for individual candidates are scrutinised for consistency and any aberrant results checked back to the actual scripts.

Reporting results

Candidates receive a Statement of Results which includes all OET results for a two-year period including the most recent test. Statements of Results include a number of inbuilt security features. Candidates also have the option of logging onto the website and viewing their results online, although these results are not official.

Staff from regulatory authorities and DIAC can apply for secure access to individual candidate result histories online.

Appeals

Because of the rigour of the assessment process as described above, the only review of results available covers statistical and identity checks.

Qualitative feedback

Candidates may seek feedback against the criteria in the Writing and Speaking sub-tests for a fee.

Re-sits

Candidates may currently re-sit sub-tests where they have not achieved the grade required by their profession. No limit on the number of times a candidate re-sits is set by the OET Centre, though specific professions may apply a limit. Professional bodies may require candidates to achieve satisfactory grades in all sub-tests at the same administration.

Development of new test materials

New test materials are developed on a regular basis for all four sub-tests. Reading and Listening materials are developed by the Language Testing Research Centre of the University of Melbourne and expert consultants. Speaking and Writing materials are developed by experienced materials writers engaged by the OET Centre and in consultation with professional educators or experienced practitioners for each profession.

Materials writing

Listening sub-test

The two listening texts are developed by health professionals according to guidelines provided by test development specialists at the University of Melbourne. In response to feedback on initial topic ideas and drafts of materials, texts are refined and then question and response formats are developed by the university team. These are subjected to trialling, analysis and further refinement. From this feedback a marking guide is developed for use by assessors.

Reading sub-test

Each Reading test consists of two articles adapted from professional journals, with multiple-choice questions. Feedback on text content is provided by health professionals and questions are developed by a team of test development specialists. Subject matter is chosen to be as inclusive as possible and to cover a range of professional interests.

The multiple-choice questions based on the texts avoid, as far as possible, drawing on candidates' specialist background knowledge.

Trialling of new Reading and Listening sub-tests

The Listening and Reading test materials are trialled on suitable subjects (health professionals with a non-English-speaking background) with due attention to test security.

Reading sub-test revision research (2009)

A research project is currently underway to enhance the validity and reliability of the Reading sub-test. New task types are being trialled with the aim of expanding the range of reading skills tested by the tasks and to increase the number of items in the test. Researchers have also canvassed the views of medical professionals on what and how they read in the modern workplace. The full report is to be delivered in mid 2009.

Reading and Listening are common to all professions. Writing and Speaking are profession-specific. For this reason, these materials are developed in consultation with the regulatory bodies and the teaching institutions for those professions.

Writing sub-test

Scenarios are developed which enable the candidate to demonstrate the written communication skills necessary to function in a professional environment. Letter content is subject to advice from specific professions regarding what may be required, such as a letter of referral or a report on patient test results.

Speaking sub-test

Scenarios are developed which enable the candidate to demonstrate spoken communication and professional listening skills. Through an iterative process involving health profession educators, OET assessors and test specialists at the OET Centre, quality of test items is assured. Interviewers are required to adopt the persona of a patient suffering a precise condition or the carer of a patient, and so present a plausible professional challenge to the candidate.

Test reliability

Reliability of the OET

The reliability of the OET sub-tests is investigated using RASCH Item Response Modelling (see McNamara 1996 for an extensive introduction). This form of analysis yields data not only on the ability of tests to discriminate across the whole cohort of test-takers, but at particular score levels of significance (especially cut-scores) for receptive skills tests (Listening and Reading). In addition, it can provide very fine analysis of the measurement quality of measures of productive skills (Speaking and Writing).

For the receptive skills, while overall reliabilities can be calculated (e.g., the overall reliability of the Listening sub-test is 0.93 in logit terms, or 0.95 (Kuder-Richardson 20) in raw score terms), it makes more sense to see how the test is discriminating at these crucial cut-score points, and here maps of the distribution of item difficulty against candidate ability enable us to check on the discriminating quality of the test.

Thus, while the overall reliability of the Reading sub-test (at or above 0.8 Kuder-Richardson 20) is somewhat lower, this is an artefact of the number of items and the generally high level of ability of the candidature; the item quality is high (all items show good discrimination or fit), and discrimination at cut-scores is high – the test is well targeted to the ability range around the cut-score.

For the productive skills, all performances are double scored, and analysed using multi-faceted RASCH measurement (McNamara 1996) using the program FACETS (Linacre 1989). This enables scores adjusted for rater severity to be produced, and provides detailed analyses of rater quality including consistency. Reports on the satisfactory quality (or otherwise) of the measurement of individual candidates are also produced, thus flagging individual performances for third rating.

A statistic called Root Mean Square Error (RMSE), the logit equivalent of the classical Standard Error of Measurement (SEM), is available from the program.

For a recent Speaking test, the RMSE for individual candidates was 0.75 logits (about one third of a score point) For the Writing test, the corresponding figure for candidates was 0.53 logits (about one quarter of a score point).

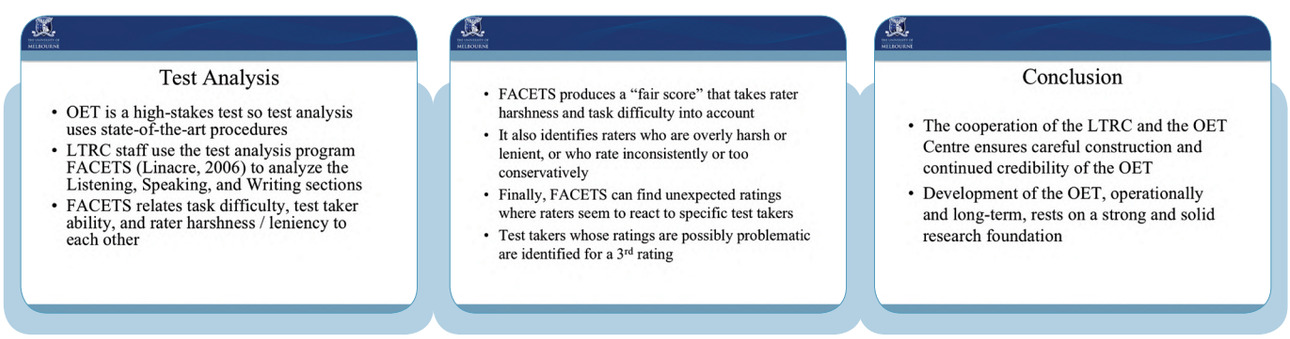
In short, this means that an extremely high quality of measurement is provided for these skills.

Feedback from RASCH analysis also assists in professional development and training for assessors.

In summary, average reliability of the individual sub-tests is:

Reading (objectively scored) 0.80
Listening (semi-objectively scored) 0.93
Speaking (subjectively scored using two raters, analysed with FACETS; fair scores reported, plus a third rating for misfitting cases) 0.95
Writing (subjectively scored as above) 0.95

Test Analysis, presentation by Dr Carsten Roever, April 2007



Validation of the OET

The validation of the OET was originally carried out at the time of its development in the late 1980s, and is the subject of reports written at that time. (McNamara 1987, 1988, 1989b)

A recommendation to develop such a test was the main finding of a report based on an earlier consultancy (Alderson et al. 1986): it recommended the creation of a new test designed to 'assess the ability of candidates to communicate effectively in the workplace'.

The test validation process is set out fully in a PhD thesis (McNamara 1990a) and in subsequent journal articles in the journal *Language Testing* (McNamara 1990b, 1991a) and elsewhere (McNamara 1989a, 1991b, 1997).

The design of the test was based on an extensive job analysis, canvassing information from the various branches of the health professions about the most common reading, listening, speaking and writing tasks required of health practitioners and their associated language demands.

Overseas trained health professionals who were now working in clinical settings were asked to assess a range of workplace communicative tasks in terms of frequency of occurrence, importance, and how difficult they found them. This analysis provided the basis for a four-skill test, with reading and listening tasks common to all professions and speaking and writing tasks mirroring the particular communicative demands of specific health professions (nursing, medicine, dentistry, physiotherapy, etc.).

Analyses of data from the test pilot and subsequent administrations in the first two years of its operation (1987-89) were then used to explore the validity of the communicative language ability construct underpinning the test design.

The test is extensively discussed as a model of testing for specific purposes in Douglas (2000). The issue of specific purpose testing in a range of contexts (including health) has been widely discussed (e.g., Jacoby & McNamara 1999; Douglas 2000; Davies 2001; Basturkmen & Elder 2004) and is seen as having two major advantages over general purpose testing.

The first advantage is the authenticity and relevance of test tasks and test-taker responses to the contexts of concern, which can enhance the validity of inferences made about a candidate's language proficiency in health-specific situations based on their test score.

The second advantage relates to the potential washback of this kind of test on the language teaching and learning activities which occur in preparation for the test.

Over the years, numerous studies have explored various aspects of the validity of the test for its specific occupational purpose.

Some of the questions posed in these studies are as follows:

- **How lifelike are the OET roleplays in their attempt to simulate interaction between a patient and health professional?** (Lumley & Brown 1996)
- **What constraints does the test situation impose on test authenticity?** (McNamara 1997)
- **What criteria do health professionals apply in judging the performance of their peers in real world contexts? Do these differ from those invoked by ESL raters when judging the quality of OET candidates' speech?** (Lumley et al. 1994; Ryan 2007)
- **Do ESL raters and doctors agree on the 'pass-fail' classification of candidates?** (Lumley et al. 1994)
- **How does a specific purpose test like the OET differ in content, format, properties and outcomes from a more general English proficiency test like IELTS?** (Elder 2007)
- **What are OET candidates' views regarding the relevance of the OET texts, topics and task types for their intended purpose?** (Elder 2007)

These past studies have all addressed the validity and reliability of the OET as a measure of specific purpose ability.

Best practice in test validation requires that validation research be conducted on an ongoing basis to ascertain that the test remains valid and relevant for its purpose and to identify areas where improvements are needed. In recent years a far-reaching program of validation research designed expressly to inform the ongoing process of refining test materials, test administration and marking procedures has been set up under a Memorandum of Understanding with the Language Testing Research Centre at the University of Melbourne.

Validation of the OET (continued)

Some of the questions addressed thus far have been:

- **What decision making processes do markers of the open-ended items on the OET listening test go though when using the official marking guide?** (Harding & Ryan 2009)
- **What features of the OET writing prompts affect the difficulty of the task for candidates?** (Knoch 2008)
- **How successfully do the OET rating criteria for speaking (e.g., intelligibility, appropriateness) discriminate between able and less able candidates?** (Chan 2006)
- **Do raters privilege some scoring criteria over others thereby giving an unfair advantage to certain candidates on the OET Speaking test?** (Iwashita & Grove 2003)
- **How stable are OET raters' judgements over time?** (Lumley & McNamara 1995)
- **How do OET raters respond to an extended program of individualised feedback on their scoring of candidates' performance on the OET speaking and writing components? Does this feedback improve the accuracy and consistency of their rating behaviour?** (Knoch 2009)
- **To what extent does the reading component of the OET encompass the diverse range of reading tasks and reading skills required to carry out their professional role?** (Elder et al. forthcoming)

The above studies draw on a range of quantitative and qualitative methodologies such as surveys, interviews, think-aloud protocols, discourse analysis, and statistical analyses of item and test score data. They attest not only to the OET Centre's concern for monitoring the validity and reliability of the Occupational English Test, but also to its concern that the test is useful and relevant to all of its stakeholders, and that it is fair for candidates.

References

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

Basturkmen, H. & Elder, C. (2004) The Practice of LSP. In A. Davies & C. Elder (eds.) *Handbook of Applied Linguistics* (pp. 673-694). London: Blackwell.

Chan, H. (2006). *Test-taker discourse and assessment criteria in the OET speaking sub-test.* Unpublished Masters Thesis. School of Languages and Linguistics, University of Melbourne.

Davies, A. (2001) The logic of testing languages for specific purposes. *Language Testing*, 18 (2), 133-147.

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

Elder, C. (2007) *OET-IELTS benchmarking study. Final Report to the Occupational English Test Centre.* 45pp. Language Testing Research Centre, University of Melbourne.

Elder, C. (2008) Comparing tests: what to tell the stakeholders? Paper delivered at the 30th Annual Language Testing Research Colloquium, Zhejiang University, Hangzhou, China, 25th-28th June.

Elder, C., Harding, L. & Knoch, U. (2009) *Reading for the health professions:* The Occupational English Reading Test revisited. Interim Report. Language Testing Research Centre, University of Melbourne.

Harding, L. & Ryan, K. (2009 forthcoming) Decision-making in marking open-ended listening test items: the case of the OET. *Spaan Working Papers in Second or Foreign Language Assessment.*

Iwashita, N. & Grove, E. (2003) A comparison of analytic and holistic scales in the context of a specific-purpose speaking test. *Prospect* 18, 3: 25-35.

Jacoby, S. & McNamara, T. (1999) Locating competence, *English for Specific Purposes* 18 (3), 213-241.

Knoch, U. (2008) Factors influencing prompt difficulty on the OET writing task for nurses. Unpublished Report. Language Testing Research Centre, University of Melbourne.

Knoch, U. (2009) Investigating the effectiveness of individualized feedback for rating behaviour: a longitudinal study. Paper presented at the 31st Language Testing Research Colloquium Denver Colorado, March 17-20.

Linacre, J. (1989) FACETS: a computer program for many-faceted Rasch measurement. Chicago: MESA Press

Lumley, T., Lynch, B. & McNamara, T. (1994) A new approach to standard-setting in language assessment. *Melbourne Papers in Language Testing* 3 (2), 19-40.

Lumley, T. & Brown, A. (1996) Specific-purpose language performance tests: task and interaction. In G. Wigglesworth & C. Elder (eds.) *The Testing Cycle: from inception to washback, Australian Review of Applied Linguistics*, Series S, Volume 13 (pp. 105-36). Canberra, Australian National University.

Lumley, T. & McNamara, T. (1995) Rater characteristics and rater bias: implications for training. *Language Testing* 12 (1), 54-71.

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

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

McNamara, T. (1989a) *The development of an English as a Second Language writing test for health professionals. Part Two of a Report to the Council on Overseas Professional Qualifications on a consultancy to develop the Occupational English Test.* University of Melbourne Department of Russian and Language Studies, Parkville Victoria.

McNamara, T. (1989b) ESP testing: general and particular. In C. Candlin & T. McNamara (eds.) *Language, learning and community* (pp. 125-42). National Centre for English Language Teaching and Research, Macquarie University.

McNamara, T. (1990a) *Assessing the second language proficiency of health professionals.* PhD Thesis. University of Melbourne.

McNamara, T. (1990b) Item Response Theory and the validation of an ESP test for health professionals. *Language Testing* 7 (1), 52-76.

McNamara, T. (1991a) Test dimensionality: IRT analysis of an ESP listening test. *Language Testing* 8 (2), 139-159.

McNamara, T. (1991b) The role of Item Response Theory in language test validation, in S. Anivan (ed.) *Current Developments in Language Testing* (pp. 165-184). SEAMEO Regional Language Centre, Singapore.

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

McNamara, T. (1997) Problematising content validity: the Occupational English Test (OET) as a measure of medical communication. *Melbourne Papers in Language Testing* 6 (1), 19-43.

Ryan, K. (2007) *Assessing the OET: the nurses' perspective.* Unpublished Masters Thesis. School of Languages and Linguistics, University of Melbourne.

OET- IELTS BENCHMARKING STUDY REPORT
EXECUTIVE SUMMARY

Associate Professor Cathie Elder
Director
Language Testing Research Centre, University of Melbourne
April 18, 2007

The report describes a small scale research study designed to compare pass thresholds on two tests currently used to determine whether overseas-trained health professionals have sufficient English proficiency to cope with studying on a clinical bridging program and, ultimately, practising their profession in Australia. The tests are the Occupational English Test (OET), designed to measure communicative competence for healthcare, and the International English Language Testing System (IELTS), designed to measure proficiency for academic purposes. While the purpose of each test is different, the two tests are currently used interchangeably by a number of professional regulatory authorities. It was therefore deemed important to determine whether the passing standards for each test were comparable and whether test takers see the tests as equally difficult and/or appropriate as measures of their English competence.

The questions addressed by the research were as follows:

1. To what extent do scores on the OET equate to scores on the IELTS (Academic)?
2. What are test takers' views regarding the difficulty and fairness of the IELTS and the OET as measures of their ability?

The report begins with an overview of research on the issue of test comparability and proceeds to describe the study itself. The tests under investigation are first compared in terms of content, format and statistical properties (including reliability) and the rather different approaches to determining pass thresholds on each are outlined. The characteristics of test takers who volunteered for the study are then presented, with particular reference to their prior experience of OET and IELTS test taking and the impact of any test preparation courses they had taken. The report then outlines methodology for the study, which involved administering both the OET and the IELTS in close succession to a group of 53 candidates and then having them complete a questionnaire canvassing their views on the relative difficulty and fairness of each test as a measure of their English language ability.

Results, based on correlational analyses and a comparison of band score/grade frequencies on each test, suggest that while the two tests tap some common features, they are not strictly equivalent in what they measure. Although the overall pass thresholds on each test are comparable, which indicates that the two tests are in this sense approximately equal in difficulty, there is a marked discrepancy in the way individuals are categorized on each measure, with some passing the OET and failing the IELTS and others doing the converse. These discrepancies can be explained in part by construct differences between the two

tests and in part by measurement error. The best fit is between an overall IELTS Band 7 and straight B grades on the OET, which are the levels currently required for a pass by most regulatory authorities. The classification agreement rate is 72%. Raising or lowering the pass thresholds increases the number of classification discrepancies across the two tests. Although there were a number of A grades on various OET components, no candidate received A grades across the board; consequently, setting the pass threshold at this higher level would have denied all candidates in the sample their professional registration. This also highlights an important difference in how final results for the two tests are obtained. In the IELTS, the overall band score is averaged across all skills components and rounded up, allowing weakness in one area to be compensated for by strength in another. The OET reports candidates' four grades separately without an overall grade; a successful candidate must therefore obtain a B grade in all four components. In the study, it was clearly easier for candidates to gain a B grade on any single skills component of the OET than to achieve a Band 7 score on the corresponding component of the IELTS; nevertheless, the across-the-board B grade requirement (compared to the average band score requirement for IELTS) means that the overall pass rates for the two tests are similar.

Questionnaire feedback from these non-native health professionals confirms what the test description and score comparisons show about the marked differences between the IELTS and the OET with respect to what they measure and how responses are elicited. More respondents rated the OET listening component as difficult than they did the IELTS listening component, giving reasons such as speed of delivery as well as the demands of taking notes while listening. Nevertheless, more candidates regarded this (OET) component as appropriate and relevant to the professional domain than they did the IELTS listening component. In contrast, more respondents evaluated the three other OET components as easy than they did the corresponding IELTS components. This was attributed to the fact that the OET texts and topics were relevant to their professional expertise and hence more accessible to them. While many candidates accept the IELTS as a valid measure of their general English proficiency, there is almost unanimous agreement among candidates that the OET is a better measure of their professional competence and hence more relevant to their needs.

It is advocated that regulatory authorities be informed about the differences between a profession-specific test and an academic English test, so that this point is better understood and choices are based on the relevance of test construct to the target ability domain, rather than views about the relative difficulty of one or other test, which, on the basis of this sample of test takers, appear unwarranted. The OET, it is argued, has far greater claims to validity than the IELTS as a measure of professional competence because of the way it has been designed and the research which underpins these design decisions. Bureaucratic imperatives demand that the pass thresholds on the two tests be comparable. For practical purposes this research has demonstrated that they are. Nevertheless, this study finds no justification for using the IELTS as the benchmark in any validation or standard-setting studies conducted by the OET Centre. Measures should instead be taken for the continuing re-

Associate Professor Cathie Elder
Director
Language Testing Research Centre, University of Melbourne

view and refinement of the OET to make it an even more robust measure and to ensure that the claims it makes about candidate performance at different grade levels are well understood by key stakeholders.

In light of the above it is recommended that OET Centre:

1. Review the OET Reading specifications with a view to increasing the number of test items and thereby boosting levels of reliability (and hence accuracy) of scores;
2. Establish clearer procedures for equating across test versions and profession-specific tasks within versions, to verify that difficulty levels are constant comparable across candidatures;
3. Commission a study to determine whether any effect for background knowledge on test performance is consistent across professions;
4. Conduct a standard-setting study involving stakeholders from the relevant professions to ensure that their views are taken into consideration in determining cut-scores between 'satisfactory' and 'unsatisfactory' performance on the various skills components of the OET.

Associate Professor Cathie Elder
Director
Language Testing Research Centre, University of Melbourne

Test Content	OET	IELTS
Skills tested	Reading Listening Speaking Writing	Reading Listening Speaking Writing
Listening	Assesses English as used in health training or clinical contexts	Assesses general English using social, education, and training contexts
	• A dialogue involving patient history taking by a health professional and a monologue in lecture format on a health-related topic; each is heard only once • Up to 20 questions in 45-50 minutes (about 80-90 scored items) • Question types: note-taking, short answer, chart completion, sentence completion, gap fill • Includes a variety of Australian accents • Responses are marked by trained listening raters. Double marking of responses close to the cut-score between B and C grades	• Monologues and dialogues between two/three people; each is heard only once • 40 items in 30 minutes • Questions types: multiple choice, short answer, notes / summary / flow chart completion, sentence completion, labelling a diagram, matching • Includes a variety of native English accents • Responses are clerically marked by a single scorer • Misspellings are not accepted
Reading	Assesses English as used in health training or clinical contexts	Assesses academic English skills
	• Two passages on health-related topics • 20-24 items in 60 minutes • Question type: multiple choice • Responses are machine-scanned	• Two reading passages appropriate for those who are entering undergraduate or graduate courses • 40 items in 60 minutes • Question types: multiple choice, sentence or prose summary completion, short-answer questions, matching lists or phrases • Responses are clerically marked by a single scorer
Speaking	Assesses contextualised interaction in a profession-specific situation	Assesses interactive speaking skills using a structured interview conducted by an examiner and recorded on a cassette tape
	• In three parts • Initial structured conversation on candidate's professional background (not assessed) • Two separate roleplays in which candidate takes his/her professional role and the interviewer acts as patient • 20-25 minutes in length	• In three parts • Part 1: introduction and interview based on familiar topics selected by the interviewer • Part 2: Individual long turn in which test-takers respond to a selected task card • Part 3: Two-way discussion linked to the Part 2 topic • 11-14 minutes in length
Writing	Assesses professional written communication	Assesses academic writing skills
	• One task • Letter of referral or a similar professional communication based on about two pages of stimulus material often in the form of case notes • 180-200 words in appropriate format • 5 minutes' reading time, then 40 minutes' writing time	• Two tasks • Task 1: Test takers review a diagram, table, or data and write about the information in their own words (20 minutes) • Task 2: Test takers write an essay on a given topic using their own experience (40 minutes) • 60 minutes in length
Total Test Time	Three hours for written sub-tests; 25 minutes for Speaking sub-tests	Three hours
Scores	• Grades for each sub-test are reported at one of five levels (A, B, C, D, E) • There is no overall OET score	• Overall band scores are reported in 0.5 increments on a 0-9 scale • Overall band score is an average of the four module scores • Listening and Reading module scores are reported in 0.5 increments • Writing and Speaking module scores are currently reported in one-point increments (to 1 July 2007)
Pass threshold set by regulatory authorities	Most regulatory authorities require a minimum B grade across all four sub-tests	Most regulatory authorities require an overall band score of 7; some also set requirements for particular module scores
Scoring procedures for Speaking	• Scoring is done retrospectively by two trained raters (and a third in problematic cases flagged by routine statistical analysis)	• Scoring is done in-country by one trained rater, the examiner who also conducts the interview
Scoring procedures for Writing	• Multi-faceted Rasch analysis creates a 'fair score' for each candidate from these scores	• Scoring is done in-country by one trained rater at the test centre
Quality control of rating	• The relative severity/leniency of each rater is calculated and scores are routinely adjusted to take account of this • Raters receive thorough initial training and are given feedback on their performance after each administration and must re-train if found to be rating inconsistently	• All Writing and Speaking raters must undergo a re-training and re-certification process every two years • Monitoring of the reliability of the Writing and Speaking responses is achieved through a sample monitoring process

Associate Professor Cathie Elder
Director
Language Testing Research Centre, University of Melbourne
April 18, 2007