



# Research and OET Development

## Ensuring the OET's Validity

**Dr Carsten Roever**

*Language Testing Research Centre*

*The University of Melbourne*

*carsten@unimelb.edu.au*

# Overview

- Development as construction and performance monitoring of a specific test form
  - Text selection
  - Paneling process
  - Test analysis
- Development also as continuing research into the OET
  - Standard setting
  - Washback
  - Equating

# Test Design

- LTRC develops two sections of the OET:
  - Listening
  - Reading
- The Speaking and Writing sections are developed by the OET Centre
- Item analysis for all sections is done at LTRC

# The Listening Section

- The listening section consists of a monologue (lecture format) on health-related topic and a dialogue involving a patient consultation by a health professional; each is heard once only
- A variety of response formats are used (note taking, summary completion, short answer, chart completion, sentence completion)

## Development: Listening Section

- A medical professional gives a 20 minute lecture on a topic
- Another medical professional does a simulated consultation with a plausible patient of about 15 minutes
- Questions are developed by LTRC staff and tried out in a paneling session
- Unclear, overly easy, or overly difficult items are cut or modified

# The Reading Section

- The Reading Section consists of two texts on medical topics
- Each text is about 800-900 words long, and currently accompanied by 10-12 multiple choice questions, which are clearly linked to paragraphs.
- Each multiple choice question has four answer options
- Questions assess comprehension of content, complex propositions, and vocabulary

## Development of the Reading Section

- LTRC developers identify 3-4 suitable texts from specialist medical journals, popular science publications, and other sources
- Texts are shortened, modified or sometimes merged to obtain two different test text
- Developers write items independently
- We often trial five answer options to facilitate later deletion / modification

- We use “logic items” (“all of the above” / “none of the above”) and negative stems (“Which statement is **not** true...”) sparingly
- Texts and items are discussed in paneling meetings
- Unclear items are cut, and the number of items per text is pared down to about 24 per text for trialing.

# Piloting

- The paneled drafts of the Reading and Listening sections are administered to between 30 and 60 paid volunteers
- The marking guide for the listening section is further refined after piloting.

## Analysis of Pilot Results

- Pilot test results are analyzed with the test analysis program QUEST, supplemented by the statistics program SPSS
- Our focus is on high overall reliability for the test and well-functioning distractors
- Reliability is the precision and consistency with which a test measures; it is a central indication of test quality
- Reliability is measured on a scale of 0 to 1, with coefficients in the range of .7 “acceptable”, .8 “good”, and .9 “excellent”
- We aim for reliability above .8, which is common for high-stakes tests

## QUEST Output for a well performing and a badly performing item:

Item 6: Infit MNSQ = 0.77  
Disc = 0.62

Categories	A [1]	B [0]	C [0]	D [0]	E [0]
Count	35	2	3	6	5
Percent (%)	68.6	3.9	5.9	11.8	9.8
Pt-Biserial	0.62	-0.14	-0.28	-0.32	-0.30
Mean Ability	0.24	-0.60	-1.05	-0.80	-0.83
StDev Ability	0.66	0.47	0.78	0.46	0.50

Item 17: Infit MNSQ = 1.30  
Disc = -.16

Categories	A [0]	B [0]	C [1]	D [0]
Count	18	22	10	2
Percent (%)	34.6	42.3	19.2	3.8
Pt-Biserial	-0.21	0.43	-0.16	-0.26
Mean Ability	-0.33	0.30	-0.37	-1.17
StDev Ability	0.70	0.71	0.70	0.54

## SPSS reliability output

Item	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
1	22.35	52.231	.387	.824
2	22.85	51.937	.388	.824
3	23.08	53.680	.282	.827
4	22.87	52.040	.382	.824
5	22.92	52.974	.260	.827
→ 6	22.46	50.489	.593	.819
7	22.87	52.707	.277	.827
8	22.65	54.231	.030	.833
9	22.42	52.955	.232	.828
10	22.54	51.704	.387	.824
11	22.42	50.916	.549	.820
12	22.92	55.445	-.148	.836
13	22.56	51.428	.423	.823
14	22.48	54.451	.004	.834
15	22.67	53.009	.197	.829
16	22.48	53.000	.211	.828
→ 17	22.94	55.859	-.220	.837

## Assembly of final version

- Items that do not improve reliability are deleted (10-12 per text), while maintaining 1-3 items per paragraph for final version.
- Unattractive distractors are modified
- Reliability in the operational runs tends to be lower because groups are more homogenous
- LTRC staff discuss the marking guide for the Listening section with the OET Centre

# Operational administration

- The OET is administered and scored by the OET Centre
- Reading responses are scanned electronically, Listening is scored by one rater, Speaking & Writing by two raters
- Results of scoring sessions (objective and rater scoring) are sent to LTRC

# Test Analysis

- OET is a high-stakes test so test analysis uses state-of-the-art procedures
- Speaking and Writing are particularly critical because they are scored by human raters, and are used to set cut scores.
- LTRC staff use the test analysis program FACETS (Linacre, 2006) to analyze the Listening, Speaking, and Writing sections
- FACETS relates task difficulty, test taker ability, and rater harshness / leniency to each other

- This allows it to adjust test takers' scores for the effect of an overly harsh or lenient rater
- FACETS produces a “fair score” that takes rater harshness and task difficulty into account
- It also identifies raters who are overly harsh or lenient, or who rate inconsistently or too conservatively
- These raters can then be retrained

- Finally, FACETS can find unexpected ratings where raters seem to react to specific test takers
- Test takers whose ratings are possibly problematic are identified for a 3<sup>rd</sup> rating
- After Writing & Speaking ratings are complete, the score distributions are used to set cut scores for the Listening and Reading sections
- Test takers within 2 points above and below the B/C borderline on Listening get a 2<sup>nd</sup> marking

# Final Report

- LTRC produces a final report on the OET administration with summary information about the test sections, rater performance, score-to-grade (band) conversions, and rescoring.
- We also look at possible revisions of test versions (items and or distractors within items) after operational analysis feedback.
- The OET Centre uses this information in rater (re-) training and assembly of future test forms

# OET Research

- The LTRC in cooperation with the OET Centre is responsible for conducting validity research on the OET
- We focus particularly on research to support and refine the inferences drawn from OET scores about communicative ability in the health professions

# The Benchmarking Study

- The OET-IELTS Benchmarking study compared the performance of a small group of test takers who completed both tests
- It showed that the overall pass rates were very similar (assuming pass marks of IELTS 7.0 average and OET B across the board)
- However, individual sections of the tests do not equate easily

- This is not surprising, as IELTS is not specific to health professionals
- IELTS is designed to assess academic English proficiency, rather than professional / vocational communicative ability
- OET is arguably closer to real-world language use of health professionals

## Future Research: Standard Setting

- Currently, judgments of performance on the OET speaking task are based on 5 linguistic criteria: *Overall Communicative Effectiveness, Intelligibility, Fluency, Appropriateness, and Resources of Grammar & Expression.*
- A study is being planned to reorient these judgments towards expert assessments of appropriate professional language use

- In a first step, samples of real-world communication from different health care settings will be collected
- Health care professionals will be recruited to provide their views of the interaction, participants' strengths and shortcomings
- From their comments, criteria for effective communication will be derived
- These criteria will then form the basis for new performance standards

- This practically oriented study will generate several sub-studies, investigating
  - methodologies for deriving criteria from assessments of (semi) authentic samples,
  - differences between health professionals and language professionals in assessing health communication,
  - differences between current statistically based standard setting approaches and criterion / performance based ones

## Future Research: Washback

- Washback is the effect of testing instrument on teaching and learning
- To what extent does teaching in OET preparation courses contribute to actual language learning rather than just increased test wiseness?
- Washback studies usually employ observation, syllabus analysis, questionnaires, and benchmarking tests

## Conclusion

- The cooperation of the LTRC and the OET Centre ensures careful construction and continued credibility of the OET
- Future research will improve the OET further, especially with regard to its specificity to health professions
- Development of the OET, operationally and long-term, rests on a strong and solid research foundation



THE UNIVERSITY OF  

---

MELBOURNE