How valid are domain experts' judgements of workplace communication? Implications for setting standards on the Occupational English Test (OET) Writing sub-test

Simon John Davidson

ORCID Identifier: 0000-0002-2242-1119

A thesis submitted in total fulfilment of the requirements of the degree of Doctor of Philosophy

June 2018

School of Languages and Linguistics
The University of Melbourne
Abstract

As part of the prerequisite to obtain professional registration and practice in Australia, International Medical Graduates (IMGs) must satisfy requirements that include satisfactory clinical knowledge and competent English language proficiency. However, there has been some stated apprehension that the minimum standards on tests used to assess language competency (including the Occupational English Test (OET), a specific-purpose language (LSP) test for health professionals), might be inadequate for successful workplace readiness. To better understand the validity of these concerns, this study examined the minimum standards on the OET Writing sub-test via the process of ‘standard setting’ – a procedure for drawing insights from appropriate stakeholders (in this case doctors with experience of workplace communication demands) about levels of proficiency viewed as satisfactory for a particular purpose. The study sought to determine the minimum levels of competence deemed appropriate for effective performance in the workplace, and also to understand the basis for the decisions made and how closely these corresponded to the construct of communication that the OET is designed to measure. A previous study (Manias & McNamara, 2016; Pill & McNamara, 2016) explored these issues in relation to the OET Speaking sub-test, whereas the current study focuses on writing – a thus far neglected area. The writing task on the OET is a letter of referral, based on a set of provided case notes. 18 health professionals (all with experience of working as medical educators, GPs or specialists) were recruited to participate in standard-setting workshops designed to elicit decisions about what level of performance on this task was deserving of a passing grade and why. To gain further insight into the basis for the standards set, verbal reports in the form of a think-aloud protocol (TAP) were employed. The doctors’ comments from the workshops and verbal reports were thematically coded and intercoder reliability checks were conducted.

Before new passing standards and ‘cut scores’ were calculated, a FACETS analysis (Linacre, 2017) was carried out to take into account any variation in domain experts’ ratings in terms of them being overly severe/l lenient or inconsistent. The final quantitative analysis yielded a somewhat more stringent passing standard than the
current one – mirroring the results of previous studies using the Analytic Judgement method (AJM) (e.g., Knoch et al., 2017 and Pill & McNamara, 2016). The new standards were compared with current OET cut scores and indicated a higher ‘fail’ rate with the current data set. The stricter passing standard established by domain experts in this study could be construed as backing indications and perceptions that the present benchmark is not set high enough and that some IMGs who are not yet communicatively competent are still joining Australian work environments with unsatisfactory communication skills. The qualitative analysis further investigated whether domain experts are competent to assess language proficiency separately from other professional skills (as stipulated by Australian federal government requirements). Some participants’ judgements (a minority overall) were influenced by views of candidates’ clinical competency, extending beyond the construct of communicative competence as defined by the OET. However, the qualitative findings, in the main, suggested that subject-matter experts were indeed attending to textual features related to what the OET is intended to measure. The central question of whether domain experts, without linguistic training, are well placed for setting the standards in a LSP test such as the OET was considered. Even though some participants’ judgements deviated somewhat from the current OET Writing sub-test criteria, validity evidence collected in this study verified, on the whole, that the subsequent new standards, derived from domain expert participation, were justified. The validity consequences of this study’s results for the OET Writing sub-test, and for LSP testing more generally, were reflected on using a unique argument-based validity framework posed by Knoch and Macqueen (in preparation). The practical and operational implications of the study’s findings for the OET were also considered.
Declaration

This is to certify that:

i. the thesis comprises only my original work towards the PhD;
ii. due acknowledgement has been made in the text to all other material used;
iii. the thesis is less than 100,000 words in length, exclusive of tables, figures, references and appendices.

Signed:

Simon John Davidson
Acknowledgements

I would like to thank first and foremost my co-supervisors Dr Ute Knoch and Assoc. Professor Catherine Elder for their generous guidance, wisdom, patience and encouragement. I express further gratitude to the overall ARC-funded project team, of which this thesis formed a part, for their advice and support and whose members included Assoc. Professor Robyn Woodward-Kron, Professor Tim McNamara, Assoc. Professor Eleanor Flynn, Professor Elisabeth Manias, Sharon Yahalom and Annemiek Huisman. I am also much obliged to Dr Ruth Sutherland and Kim Munro for their insights during the data coding process.

I am grateful to Cambridge Boxhill Language Assessment and Dr Barbara Ying Zhang and her team for supplying the test candidate writing samples used in this research. I would also like to thank the 18 health professionals who participated in this study for their valuable input and time. acknowledgement is given to the ARC-funded studentship for giving me financial support throughout and the University of Melbourne’s School of Language and Linguistics, particularly at the end of my candidature. I am also appreciative of the International Research Foundation for English Language Education (TIRF) for their further financial support. To those from the School of Language and Linguistics at the University of Melbourne who attended the first initial pilot workshop and also again Assoc. Professor Eleanor Flynn and Professor Elisabeth Manias who helped with the second pilot, many thanks for your feedback and assistance.

Finally, I would like to thank my family and friends who have motivated and encouraged me and especially to my partner Cathy. Love to you and our son Joe who arrived a little earlier than we expected in the last year of writing up this thesis. I couldn’t have got where I am today without your love and support.
# Contents

List of tables................................................................................................................... xi
List of figures.................................................................................................................... xi
List of abbreviations......................................................................................................... xii

## CHAPTER 1: INTRODUCTION

1.1 Background .............................................................................................................. 1
1.2 Role of communication in applying for medical registration in Australia............ 3
1.3 The Occupational English Test (OET)...................................................................... 5
1.4 The OET Writing sub-test.......................................................................................... 7
1.5 The impetus for the current research......................................................................... 9
1.6 The study and research questions............................................................................. 11
1.7 Outline of the thesis................................................................................................. 14

## CHAPTER 2: LITERATURE REVIEW

2.1 Introduction.............................................................................................................. 17
2.2 International medical graduates (IMGs)................................................................. 18
2.2.1 Workforce migration and globalization.............................................................. 18
2.2.2 Perceptions of IMGs in the Australian context................................................ 20
2.2.3 Registration procedures and requirements....................................................... 23
2.3 Testing language readiness for healthcare in the Australian context.................. 25
2.4 Language for specific purposes (LSP) testing......................................................... 26
2.4.1 Communicative competence................................................................................. 26
2.4.2 Theoretical justification for LSP testing............................................................. 30
2.4.3 Debates about LSP testing.................................................................................. 33
2.4.3.1 Specificity...................................................................................................... 33
2.4.3.2 Inseparability................................................................................................ 35
2.4.3.3 Authenticity.................................................................................................. 37
2.4.4 Domain expert involvement in LSP research.................................................... 38
2.4.5 LSP testing and the OET.................................................................................... 42
2.5 Standard setting...................................................................................................... 46
2.5.1 Background......................................................................................................... 47
2.5.2 Categorization of standard-setting methods...................................................... 50
2.5.3 Selecting a method............................................................................................ 51
2.5.4 Evaluation of different methods......................................................................... 53
2.5.4.1 The Contrasting Groups and Borderline Group methods.......................... 55
2.5.4.2 The Body of Work method.......................................................................... 56
2.5.4.3 The Analytic or Analytical Judgement method.......................................... 58
2.5.4.4 The Performance Profile method.............................................................. 59
2.5.4.5 The Examinee Paper Selection method...................................................... 60
2.5.4.6 The Angoff method..................................................................................... 61
2.5.4.7 The Bookmark method.............................................................................. 62
2.5.5 Studies and reports on standard setting............................................................. 64
2.5.5.1 Language testing and standard setting....................................................... 64
2.5.5.2 Health professions and standard setting.................................................... 65
2.5.5.3 Standard-setting research on the OET....................................................... 69
2.5.6 Issues in standard-setting research.................................................................... 72
3.6.4.3 Writing response allocation to participants.................133
3.6.4.4 Workshop procedures........................................135
    3.6.4.4.1 Training workshops......................................135
    3.6.4.4.2 TAP sessions.............................................136
3.6.4.5 Data analysis procedures....................................138
    3.6.4.5.1 Qualitative analysis procedures........................138
        3.6.4.5.1.1 Applying codes...................................140
        3.6.4.5.1.2 Coding validation.................................143
        3.6.4.5.1.3 Standard-setting evaluation......................146
    3.6.4.5.2 Quantitative analysis procedures.......................147
        3.6.4.5.2.1 Calculation of cut scores.......................147
        3.6.4.5.2.2 Judge behaviour..................................148
3.7 Argument-based validity framework and sources of evidence from
      this study....................................................................149
3.8 Chapter summary..............................................................152
CHAPTER 4: RESULTS............................................................153
4.1 Introduction........................................................................153
4.2 Qualitative results – workshops/TAPs ..................................153
    4.2.1 Performance level judgement..................................155
    4.2.2 Task fulfilment....................................................155
    4.2.3 Content..............................................................157
    4.2.4 Organisation........................................................162
    4.2.5 Expression..........................................................165
    4.2.6 Presentation.........................................................168
    4.2.7 Professionalism....................................................171
    4.2.8 Audience recognition.............................................172
    4.2.9 Other.........................................................................175
    4.2.10 Performance level judgements and decision-making........176
    4.2.11 Performance level judgements and variability................180
        4.2.11.1 Patient identification....................................180
            4.2.11.1.1 Cambridge Boxhill Language Assessment
                        and patient identification........................186
                4.2.11.2 Organisation...........................................187
                4.2.11.3 Expression.............................................189
                4.2.11.4 Presentation...........................................190
        4.2.13 Summary..........................................................191
4.3 Qualitative results – final evaluations................................192
    4.3.1 Training workshop evaluation..................................193
    4.3.2 Final evaluation of the standard-setting procedure........194
    4.3.3 TAP session evaluation..........................................197
    4.3.4 Summary..................................................................198
4.4 Quantitative results..........................................................198
    4.4.1 FACETS analysis....................................................199
        4.4.1.1 Rater severity/leniency....................................200
        4.4.1.2 Rater consistency..........................................205
    4.4.2 New cut scores.......................................................205
    4.4.3 Summary..............................................................208
CHAPTER 5: DISCUSSION

5.1 Introduction
5.2 Domain expert decision-making process during standard setting
5.3 Construct relevance of domain experts’ judgements
  5.3.1 Construct relevance
  5.3.2 Construct disparity
    5.3.2.1 Clinical competence
    5.3.2.2 Patient identification
    5.3.2.3 Audience recognition
    5.3.2.4 Test candidate background
    5.3.2.5 The OET Writing sub-test task and authenticity
5.4 Variation in qualitative data
5.5 Variation in quantitative data
5.6 Domain expert feedback
5.7 New cut scores
5.8 Chapter summary

CHAPTER 6: CONCLUSION

6.1 Introduction
6.2 Summary of the findings around an argument-based approach to LSP test validity
6.3 Implications
6.4 Contribution of the study
6.5 Limitations
  6.5.1 Sampling of participants
  6.5.2 Sampling of scripts
  6.5.3 Comparative cut scores
  6.5.4 Standard-setting procedures
6.6 Future research areas
6.7 Summary

REFERENCES

APPENDICES

Appendix A – Test prompt and writing sample, Task 1 ‘Brian Edwards’ (T1-BE)
Appendix B – Test prompt, Task 2 ‘Betty Johnson’ (T2-BJ)
Appendix C – AJM training response form
Appendix D – AJM take-home response form (7 categories)
Appendix E – Alternate AJM response form (12 categories)
Appendix F – Performance Profile method response form
Appendix G – ‘Think-Aloud’ pre-task form
Appendix H – ‘Think-Aloud’ response form and guidelines
Appendix I – Workshop attendees by medical practitioner specialisation, current/recent workplace(s) and role(s), years of experience since registration in Australia, and years of experience in a supervisory role
Appendix J – Workshop attendees by supervision of an overseas-trained medical practitioner, years/months of overseas-trained medical practitioner supervision, supervision of non-native English speaker/s
Appendix K – Workshop level of workplace interaction with non-native English-speaking colleagues
Appendix L – Number of referral letters (or similar pieces of communication) written by a non-native English-speaking medical practitioner read per month
Appendix M – Workshop attendees by country of birth, native language/s, country where majority of medical training was undertaken
Appendix N – Plain Language Statement (PLS) and Consent form
Appendix O – Participant pre-task form
Appendix P – Participant background information form
Appendix Q – Training evaluation form
Appendix R – Final evaluation form
Appendix S – Think-aloud final evaluation form
Appendix T – Task 1 ‘Brian Edwards’, T1-BE-74 test candidate sample
Appendix U – Data dictionary
List of tables

Table 1.1: OET Writing sub-test conversion to band scores ........................................9
Table 2.1: Decisions inference based on an LSP test (Knoch & Macqueen, in preparation) ................................................................. 92
Table 3.1: Code/theme overlap ........................................................................ 144
Table 3.2: Argument-based approach and sources of evidence from this study .... 151
Table 4.1: Themes and sub-themes ................................................................. 154
Table 4.2: Combined codes under ‘Case Note Content’ (CNC) .......................... 159
Table 4.3: Further examples of ‘Case Note Content’ (CNC) ............................... 160
Table 4.4: Inclusions of ‘Patient Identification’ (PID) ...................................... 181
Table 4.5: Rater measurement report .............................................................. 201
Table 4.6: New vs current OET cut scores ...................................................... 207
Table 4.7: ‘Pass’ vs ‘fail’ rates with former and new cut scores applied .......... 207
Table 5.1: Themes and sub-themes valued by domain experts ......................... 213
Table 5.2: Argument-based validity framework and backing from this study .... 236

List of figures

Figure 2.1: Building blocks of an interpretive argument (adapted from Kane (1992) in Knoch & Elder (2013)) ............................................. 81
Figure 4.1: ‘Variable’ or ‘Wright’ map ................................................................ 199
Figure 5.1: Domain expert decision-making process during standard setting .... 214
<table>
<thead>
<tr>
<th>List of abbreviations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ACTFL</strong> – American Council on the Teaching of Foreign Languages</td>
</tr>
<tr>
<td><strong>AJM</strong> – Analytic or Analytical Judgement method</td>
</tr>
<tr>
<td><strong>AMC</strong> – Australian Medical Council</td>
</tr>
<tr>
<td><strong>AHPRA</strong> – Australian Health Practitioner Registration Agency</td>
</tr>
<tr>
<td><strong>ARC</strong> – Australian Research Council</td>
</tr>
<tr>
<td><strong>AWOS</strong> – areas of workforce shortage</td>
</tr>
<tr>
<td><strong>ASLPR</strong> – Australian Second Language Proficiency Rating</td>
</tr>
<tr>
<td><strong>BOW</strong> – Body of Work method</td>
</tr>
<tr>
<td><strong>CEFRI</strong> – Common European Framework of Reference</td>
</tr>
<tr>
<td><strong>CELBAN</strong> – Canadian English Language Benchmark Assessment for Nurses</td>
</tr>
<tr>
<td><strong>CLB</strong> – Canadian Language Benchmarks</td>
</tr>
<tr>
<td><strong>COPQ</strong> – Council on Overseas Professional Qualifications</td>
</tr>
<tr>
<td><strong>EAL</strong> – English as an additional language</td>
</tr>
<tr>
<td><strong>ESL</strong> – English as a second language</td>
</tr>
<tr>
<td><strong>EFL</strong> – English as a foreign language</td>
</tr>
<tr>
<td><strong>ESB</strong> – English-speaking backgrounds</td>
</tr>
<tr>
<td><strong>ESP</strong> – English for specific purposes</td>
</tr>
<tr>
<td><strong>GMC</strong> – General Medical Council (UK)</td>
</tr>
<tr>
<td><strong>HP</strong> – health professional</td>
</tr>
<tr>
<td><strong>iBT</strong> – internet-based test (TOEFL)</td>
</tr>
<tr>
<td><strong>IELTS</strong> – International English Language Testing System</td>
</tr>
<tr>
<td><strong>JQC</strong> – just qualified candidate</td>
</tr>
<tr>
<td><strong>IMG</strong> – international medical graduate</td>
</tr>
<tr>
<td><strong>INMG</strong> – international nurse and midwife graduate</td>
</tr>
<tr>
<td><strong>IRT</strong> – item response theory</td>
</tr>
<tr>
<td><strong>LSP</strong> – language for specific purposes</td>
</tr>
<tr>
<td><strong>MBA</strong> – Medical Board of Australia</td>
</tr>
<tr>
<td><strong>MCC</strong> – minimally competent candidate</td>
</tr>
<tr>
<td><strong>MCQ</strong> – multiple-choice question</td>
</tr>
<tr>
<td><strong>NCSBN</strong> – National Council of State Boards of Nursing (USA)</td>
</tr>
<tr>
<td><strong>NESB</strong> – non-English-speaking background</td>
</tr>
<tr>
<td><strong>OET</strong> – Occupational English Test</td>
</tr>
<tr>
<td><strong>OIB</strong> – ordered item booklet</td>
</tr>
<tr>
<td><strong>OPB</strong> – ordered profile booklet</td>
</tr>
<tr>
<td><strong>OSCE</strong> – objective structured clinical examination</td>
</tr>
<tr>
<td><strong>OTD</strong> – overseas-trained doctor</td>
</tr>
<tr>
<td><strong>PLAB</strong> – Professional and Linguistic Assessment Board (UK)</td>
</tr>
<tr>
<td><strong>SMCC</strong> – simulated minimally competent candidate</td>
</tr>
<tr>
<td><strong>TLU</strong> – target language use</td>
</tr>
<tr>
<td><strong>TOEFL</strong> – Test of English as a Foreign Language</td>
</tr>
<tr>
<td><strong>TOEIC</strong> – Test of English for International Communication</td>
</tr>
<tr>
<td><strong>TRAB</strong> – Temporary Registration Assessment Board (UK)</td>
</tr>
</tbody>
</table>
Chapter 1: Introduction

This chapter presents an overview of the thesis. First, the key issues and context of the study are provided. This includes a depiction of the significance of communication in healthcare and an explanation of the role of communication in the professional registration process in Australia. Also, a description of the Occupational English Test (OET) and in particular the Writing sub-test, which is central to this study, is given. Third, the motivation for the current study is specified. Then, the study’s aim and research questions are presented. Finally, an outline at the end of this introduction chapter offers a preview of the remainder of this thesis.

1.1 Background

Being able to communicate effectively is an important attribute in any work environment and particularly so in health due to patient care and safety concerns. Written correspondence is an important facet of healthcare communication (Keely, Myers, Dojeiji, & Campbell, 2007) as a high standard is crucial in successful patient diagnoses and handover. If inter and intraprofessional written communication is not at an acceptable level, patient treatment may be impinged upon and peoples’ lives could be put at risk (Manias, Jorm, & White, 2008; Tjia, Mazor, Field, Meterko, Spenard, & Gurwitz, 2009). Effective and clear written communication is therefore a vital skill for all health professionals (HPs).

At times, over previous decades, Australia has faced a lack of doctors with workforce shortages being addressed, in part, through the recruitment of overseas-trained HPs to fill the gap – HPs with qualifications from abroad include International Medical Graduates (IMG’s) (Australian Government Department of Health and Ageing, 2008, 2011; Barton, Hawthorne, Singh, & Little, 2003; Hawthorne, 2012). IMGs are defined as doctors who seek employment in a country which is different from that in which they received their initial training and work experience and where the medium of workplace communication is not their mother tongue (Royal Australasian College of
Physicians (RACP), 2017). Almost 40% of medical practitioners registered in Australia fit within this definition and the healthcare system is likely to be dependent (to some extent) on such doctors for some time (Carver, 2008; Hawthorne, 2012). Doctor shortages are apparent in general practice and particular areas of medical specialty and are especially evident in remote and rural parts of Australia (Hawthorne, 2012).

Nevertheless, concerns have been raised about IMG’s ability to fulfill their role effectively in the workplace. There may be a view among some doctors and members of the public that IMG’s communication skills are ‘problematic’ even after they have fulfilled all the requirements of Australian medical registration (see e.g., Elkin, Spittal & Studdert, 2012; Harding, Parajuli, Johnston, & Pilotto, 2010; Louis, Lalonde, & Esses, 2010; Pill, 2013). A number of studies have investigated some of the general concerns (e.g., language and cultural issues) that IMGs may encounter in English-speaking healthcare situations (see e.g., Baker & Robson, 2012; Dorgan, Lang, Floyd, & Kemp, 2009; Hall, Keely, Dojeiji, Byszewski, & Marks, 2004). Many studies have focused on spoken communication in healthcare environments (see e.g., Chur-Hansen, 1997; Couser, 2007; Crawford & Candlin, 2013; Duncan & Gilbey, 2007; McDonnell & Usherwood, 2008; Pilotto, Duncan, & Anderson-Wurf, 2007; Saxena, Dennis, Vagholkar, & Zwar, 2006). Communication skills training (e.g. consultation skills) is often implemented at various stages according to university coursework or workplace requirements to ensure IMG’s have acceptable standards of safety and quality, however hardly any current IMG training courses take into account the written aspect needed (Dorgan et al., 2009; Francois, 2012; Konno, 2006; Woodward-Kron, Stevens, & Flynn, 2011). Hence, IMG’s written communication in the work environment may be viewed as less than satisfactory. Furthermore, few studies explicitly examine IMGs English writing proficiency in healthcare contexts and written communication is a neglected area both in medical training and research (Hall et al, 2004; Woodward-Kron et al., 2011). The written communication aspect in healthcare settings is the focus of this thesis.
1.2 Role of communication in applying for medical registration in Australia

IMGs who intend to register and practice in a country other than the one in which they were originally trained are required to prove their professional workplace knowledge and competence as well as separately demonstrate their language proficiency. Two major tests are currently recognized in Australia for English-language proficiency screening for health professionals:

1) The Occupational English Test (OET) – overall pass with a minimum grade of B (the test is scored on a scale of A-E in each sub-test, R, W, S, L)
2) The International English Language Testing System (IELTS) academic test – minimum score of 7 (the test is scored on a scale of 1 – 9 in each sub-test, R, W, S, L)

Whereas the IELTS is a general-purpose language test used for a range of different purposes, the Occupational English Test (OET), which is the particular subject of enquiry in this study, is a specific-purpose language (LSP) test designed expressly for the healthcare context. An LSP test is a distinct form of communicative language testing that differs from a general-purpose language test such as IELTS by including subject-specific content (described in detail in literature review section 2.4.)

Since 2010, a number of changes were made as to how IMGs can attain medical registration in Australia. These may have had an impact on how the language tests used to assess English-language proficiency for workplace readiness were perceived by stakeholders. State-based regulatory bodies for the health professions were replaced by a national accreditation and registration scheme in July 2010. The new scheme involved new legislation and harmonisation of procedures and standards which had previously been different across states. Ten groups of health professions were initially regulated under the new national system which was a significant change and had been suggested for some time (see e.g., Birrell & Schwartz, 2006; McGrath, 2004; Spike, 2006). Four additional professions were added in July 2012. The new
national board developed five mandatory registration standards including English-language skills.

As the new national registration scheme’s processes were applied, a degree of criticism and misunderstanding was noted by stakeholders. The change to the new system was one of the triggers for the Federal Minister for Health and Ageing to initiate an inquiry in November 2010. At this inquiry, further disapproval was expressed in relation to the newly implemented procedure which resulted in a report: ‘Lost in the Labyrinth: Report on the inquiry into registration processes and support for overseas trained doctors’ (House of Representatives Standing Committee on Health and Ageing, 2012). Submissions to the inquiry included a review of the language tests used and of the existing language proficiency standard. The report made several recommendations relating to language assessment, including a reappraisal of the present language skills standard itself.

Low standards of communication in the workplace are often attributed to unduly low passing standards on language tests like the OET used as a hurdle requirement for IMGs applying to practice in Australia. Indeed, the appropriateness of these passing standards have been questioned (e.g., Birrell & Schwartz, 2006). Some have argued that a test such as the OET does not perform 100% successfully in its ‘gate-keeper’ role (Lumley, 1995, 1998). In other words, there are concerns that IMGs from a non-English-speaking background (NESB) may pass the test, but still struggle with clinical interactions and be seen by colleagues and supervisors to not have the required English-language proficiency to perform adequately and safely in the workplace. Tests like the OET are ‘high-stakes’ in the sense that they “bear a heavy burden of responsibility for making decisions affecting the careers of candidates and the people for whom they are subsequently responsible” (Elder, Pill, Woodward–Kron, McNamara, Manias, Webb, & McColl, 2012, p. 410). Therefore, close scrutiny of the tests and associated decisions about who is fit to practice are in order. Ensuring that passing standards are robust is all the more important in the case of workplace writing skills where few opportunities exist for IMG training. The OET and the Writing sub-test
in particular is the focal point of this thesis. The OET is described in more detail in the following sections.

1.3 The Occupational English Test (OET)

The Occupational English Test (OET) is a language for specific purposes (LSP), ‘high-stakes’ test. The test measures the English-language proficiency of IMGs wanting to practice in an English-speaking context. The OET is conducted in more than 40 countries around the world in 12 health-related occupations: dentistry, dietetics, nursing, occupational therapy, optometry, pharmacy, physiotherapy, podiatry, radiography, speech pathology, veterinary science and medicine, which is the focus of the current study. Currently, OET results are accepted as a gauge of English-language proficiency by medical registration authorities in Australia, New Zealand, Singapore, Dubai, Namibia, the UK and Ireland (OET, 2017). Under Australian federal law, in order for IMGs to be registered, professional knowledge/competence and English-language proficiency are assessed separately (McNamara, 1996).

The OET test was originally designed to assess candidates’ capability to practise under supervision and contribute to bridging programs. In the late 1980s, a panel of language assessors reported on the form of the OET and proposed that the test should assess “the ability of candidates to communicate effectively in the workplace” (Alderson, Candlin, Clapham, Martin, & Weir, 1986, p. 3). Under contract from the Australian government, further research, development and modification of the test was carried out by McNamara (1990, 1996). The OET test was first owned by the National Languages and Literacy Institute of Australia and then passed to the Centre of Adult Education, Melbourne, which later joined with the Boxhill Institute for Technical and Further Education. Ownership of the test was acquired by Cambridge Boxhill Language Assessment in 2013, a collaborative endeavor between Cambridge English Language Assessment and Box Hill Institute.

The OET test content includes material of a general health-related and medical nature. There are four sub-tests: Writing, Speaking, Listening and Reading. The same Listening
and Reading sub-tests are undertaken by all test candidates. The Writing and Speaking sub-tests on the other hand, are specific to each occupation. For example, test materials for nurses are different to those for doctors. Assessment of the Writing and Speaking sub-tests are conducted by trained assessors with a background in teaching English as a second/foreign language (ESL/EFL). Final scores for each sub-test are converted to grades from A (the highest level) to E. Each sub-test is assessed separately and there is no overall grade such as in other high-stakes tests like the International English Language Testing System (IELTS). According to the descriptors published on the OET website, a ‘B grade’ demonstrates a “high level of performance (i.e. ability to use English with fluency and accuracy adequate for professional needs)” (OET, 2017). A ‘B grade’ in each of the four skill areas (Writing, Speaking, Listening and Reading) is considered a ‘pass’ for English-language proficiency hurdles by medical registration boards in countries where the OET is accepted. These grade descriptors were however devised by the original test designers and the determination of the minimum standard was not based on any formal standard-setting procedure (McNamara, 1990). The original passing scores were aligned with the Australian Second Language Proficiency Rating Scale (ASPLR) (Ingram, 1984) at ASLPR 3 (nine levels from lowest to highest: 0, 0+, 1-, 1, 1+, 2, 3, 4, 5) by the then Australian Government Council for Overseas Professional Qualifications (COPQ). McNamara (1990) found that that it was problematic to set a pass score equivalent to ASLPR 3 due to three reasons: 1) considerable variability both between and within the six doctor raters (all from overseas qualified non-native English speaking backgrounds); 2) significant “overlap with scores allocated to candidates at surrounding ASLPR levels; and 3) “there was some doubt about the reliability of the ASLPR ratings themselves” (p. 373-374). Since then, Pill and McNamara (2016) further note that “the cut scores in current use... are well established, although their provenance is unclear” (p. 7). How Cambridge Boxhill Language Assessment have updated and maintained the current scores and standards is not known or publically accessible for possible commercial and security reasons. Thus, this study attempts to shed light on the OET Writing sub-test’s performance standards.
1.4 The OET Writing sub-test

In the OET Writing sub-test, candidates are provided with a set of case notes from which to compose a letter of referral or other professionally relevant text. Largely, the letter is one of referral, however, for some professions, it may be a different type of letter task such as a letter of transfer or discharge, a letter to advise or inform a patient or carer and variations such as replying to a complaint (OET, 2017). Test takers are provided with an exact task, including the audience they are writing to. They must draw on case notes in their response and include information about treatment, in addition to the issues that need to be dealt with by the other health professional (receiver of the letter). Test candidates are required to handwrite approximately 180-200 words within a time frame of 45 minutes, which includes 5 minutes’ reading time (OET, 2017). Sample test prompts and writing samples (see Appendix A for an example) are provided for test candidates on the OET website. In regard to letter format and layout the OET website states:

A number of different formats are in accepted use by health professionals in different local contexts. There is therefore no single particular format that you have to use in your response in the OET Writing sub-test. It is important that your letter is clearly laid out and appropriate for the particular task, but there is no set OET layout that you have to use (OET, 2017).

However, concerning letter format/layout for doctors, other somewhat contradictory materials can also be accessed online such as from the Royal Australian College of General Practitioners (RACGP) which clearly states their preferred format for referral letters like those used in the OET Writing sub-test. RACGP (2016) has produced a handbook: ‘Referring to other medical specialists: A guide for ensuring good referral outcomes for your patients’. The guide presents an overview of accepted good practice with the aim of providing support for doctors in writing quality referral letters to other medical practitioners. Therefore, while the OET states there are a variety of referral letter formats that are accepted in the simulated test context, medical health professionals themselves, via the RACGP in a real-world situation, favour a particular letter format.
The OET Writing sub-test is currently assessed using a set of five criteria: ‘Overall Task Fulfilment’ – including whether the response is of the required length; ‘Appropriateness of Language’ – including the use of appropriate vocabulary and tone in the response, and whether it is organized appropriately; ‘Comprehension of Stimulus’ – including whether the response shows the situation has been understood and relevant rather than unnecessary information has been provided to the reader; ‘Control of Linguistic Features (Grammar and Cohesion)’ – including how effectively grammatical structures and cohesive devices of English have been used; ‘Control of Presentation Features (Spelling, Punctuation and Layout)’ – including how these aspects affect the message that has been communicated. The writing responses that test candidates produce are graded against these five criteria with each measure having six levels, 1 – 6. Level 6 signifies the highest response (OET, 2017). Cambridge Boxhill Language Assessment’s full rating scale is confidential and consequently cannot be disclosed.

The writing performances are rated independently by two individual assessors against the five criteria and without prior knowledge of a candidate’s performance on the other sub-tests. The Writing sub-test uses the statistical procedure, many-facet Rasch analysis (McNamara 1996), and FACETS software (Linacre, 2017) to arrive at an averaged or ‘fair score’ that evens out the relative severity or leniency of the test raters on each occasion the test is conducted (explained in detail in the method chapter). The fair score (range 1 – 6) is then converted into one of five grades (A – E) and corresponding band scores as shown below in table 1.1 (OET, 2009). The band score to grade conversion remains constant across test versions. The decision that a score of 4.8 or above (B grade) can be considered adequate for professional purposes is not based on any known or publically available formal standard-setting procedure.
<table>
<thead>
<tr>
<th>Band</th>
<th>Score Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Band A</td>
<td>5.6 and above</td>
</tr>
<tr>
<td>Band B</td>
<td>4.8 – 5.5</td>
</tr>
<tr>
<td>Band C</td>
<td>4.2 – 4.7</td>
</tr>
<tr>
<td>Band D</td>
<td>3.4 – 4.1</td>
</tr>
<tr>
<td>Band E</td>
<td>3.3 and below</td>
</tr>
</tbody>
</table>

Table 1.1: OET Writing sub-test conversion to band scores

The OET test construct is embodied as professional communication in health and medical contexts and therefore to strengthen the validity of the OET Writing sub-test, ongoing validation studies need to be carried out. McNamara (2000) acknowledges that with the development of communication in healthcare domains and because of institutional, technological, socio-political, and other evolutions, test tasks should be updated and redeveloped in line with these changes across professions. This is a vital part of ensuring test validity that should be undertaken regularly. Changes to test tasks may require modifications to both the rating scale and the passing standards, as discussed further below.

1.5 The impetus for the current research

A number of concerns in regard to the ability of a language for specific purposes (LSP) test such as the OET to adequately assess the written English-language proficiency of IMGs for workplace readiness have been noted (described in detail in literature review section 2.4). Firstly, there is a tendency to blame this possible inadequacy on the test itself as it may be regarded as not capturing what is important to the actual healthcare workplace domain (Lumley, 1998). Another consideration is whether the passing standard on the OET has been set too low. Issues such as these were the incentive for two research projects conducted at the University of Melbourne. Both projects were funded through separate Linkage grants won from the Australian Research Council (ARC) by a team of researchers at the University of Melbourne with Cambridge Boxhill Language Assessment as Industry Partner. The first project, ‘Towards improved
healthcare communication: Development and validation of language proficiency standards for non-native English-speaking health professionals’ (LP0991153) focused on the OET Speaking sub-test (Elder, McNamara, Woodward-Kron, Manias, Webb, McColl, & Pill 2013). The second project, ‘Towards improved quality of written patient records: Development and validation of language proficiency standards for writing for non-native English-speaking health professionals’ (LP130100171), of which this thesis forms a part, focussed on the OET Writing sub-test (Knoch, Elder, Woodward-Kron, Flynn, Manias, McNamara, Zhang & Huisman, 2017). The overall project was conducted by an interdisciplinary team of researchers from the School of Languages and Linguistics, Medical Education Unit and Melbourne Medical School at the University of Melbourne, and the Department of Nursing at Deakin University. This undertaking was a major part of an ongoing program of validation research on the OET. The project had three stages:

- Phase 1: Enquiry into health professionals’ (doctors and nurses) views of what constitutes effective writing in healthcare settings;
- Phase 2: Using the findings of phase 1 to establish more professionally relevant criteria for assessing performance on the writing component of the OET;
- Phase 3: Setting minimum standards involving health professionals in determining the minimum standards of writing performance on the OET.

While the project as a whole was concerned with two professions: medicine and nursing, the current study, undertaken independently by this researcher, but as part of this overall project, focusses on medicine only and has furnished evidence in relation to phase three of the wider project, setting minimum standards. This has been achieved through a standard-setting procedure that involved focus group discussions and verbal protocols with a sample of doctors with training and/or work experience in Australia who were asked to judge the minimum standard of written English-language competence for professional readiness.

As noted previously, if robust English-language standards are necessary for ensuring IMG’s workplace readiness, this leads to the question concerning who should be
making the decisions about fitness for practice. This is a critical issue for LSP tests which are used to assess performance in specialist domains of use. In language testing contexts, a number of studies have shown significant divergence in judgements between the language experts who traditionally assess performance on LSP tests and those with experience of the relevant professional context (see e.g., Brown, 1993; Elder, 1993; Knoch, 2014). The judgements of such professionals and the criteria which underpin them should not be ignored if a test is to be considered fit for its specific purpose. Douglas (2001a) and Hamp-Lyons and Lumley (2001) state that the recognition of professionally relevant criteria is one of the most necessary concerns for research and practice in LSP. The need for professionally relevant criteria for scoring performance is necessary, but there is also the necessity for professional input into the standard-setting process (i.e. how much language proficiency is enough for safe and effective workplace practice). This is because appropriate decisions should be made about test candidates on aspects of performance that are relevant to the professional domain.

1.6 The study and research questions

The process of setting passing standards is known as ‘standard setting’ and is a vitally significant step in the procedure of guaranteeing that defensible inferences are made from test scores. It requires adherence to a methodical practice to arrive at agreement on whether a candidate has reached a given standard by deciding a ‘cut score’. A cut score is a place on a test’s score range that is used to establish if a certain assessment score is adequate or not for some purpose (Zieky, Perie, & Livingston, 2008); whether this be to embark on training, to exit from a particular course or to enter the workplace. For example, test takers on the OET with scores at or above a specified cut score are deemed linguistically proficient/competent for the health professional workplace. They will be allowed to practice their profession in Australia (subject to also demonstrating adequate levels of clinical competence via a subsequent assessment procedure).
Setting a language proficiency cut score and passing performance standard level, via a carefully designed standard-setting procedure and with the input of health professionals themselves, is essential to be sure that test candidates are proficient enough to communicate effectively in the professional domain. As noted, it is generally accepted in LSP testing that in the process of setting standards domain experts should be involved as they have the understanding and experience of communication requirements of the professional realm. However, whether domain experts are well positioned (as non-language experts) to determine valid standards should be considered as they may attend to features of performance that are not related to communication per se. Questions may remain about the validity of subject-matter experts’ judgements which have seldom been explored.

Ultimately, this study aims to answer the question of whether domain experts, without language training, are equipped for setting valid standards on a specific-purpose language (LSP) test such as the OET. To what extent are health professionals, without linguistic training, able to attend to issues of language and communication independently of clinical competence which is their area of expertise? In addition, can domain experts agree on what matters for effective communication? Furthermore, do the samples of language and communication elicited outside the workplace setting allow sufficient opportunities for them to make these judgements and feel confident about them? Affirmative answers to these questions are needed if we are to give credence to their judgements. These considerations are further explored in this thesis.

The investigation described in this thesis aims to set minimum standards of performance on the OET Writing sub-test via the process of ‘standard setting’ by eliciting insights from domain experts, in this case doctors, about the level of proficiency deemed sufficient for safe professional practice. Following on from a previous investigation of standard setting in relation to the Speaking component of the OET test (described in the literature review), the present study utilizes a specific standard-setting method (Analytic Judgement Method (AJM)) to arrive at such minimum standards. The reasons for using this particular standard-setting procedure and the results of pilot studies are described in the method chapter. A thematic
analysis of panel discussions and verbal reports (think-aloud protocols (TAPS)) is carried out to explore the basis of doctors’ judgements about suitable cut scores and thereby to illuminate the standard-setting process. This type of qualitative approach is an underutilized aspect of standard-setting studies.

More specifically, the study reported in this thesis, ‘How valid are domain experts’ judgements of workplace communication? Implications for setting standards on the Occupational English Test (OET) Writing sub-test, seeks to answer four research questions:

1. What criteria do standard-setting participants use as a basis for their decisions in judging writing responses and to what extent are these decisions language-based?
2. Is there any variability between judges in what they attend to while setting standards?
3. How do standard-setting judges view the process and outcome of the standard-setting procedure?
4. What occupational specific standards (cut scores) do doctors set on the Occupational English Test (OET) Writing sub-test?

This study draws on both qualitative and quantitative data. In order to answer research questions one and two, doctors’ assessments of OET test candidate writing samples are analyzed qualitatively to establish which aspects of performance inform their judgements. Research question three further investigates any variation between participants. Panellists’ views of the standard-setting procedure are also captured in their completion of final evaluation questionnaires. A qualitative analysis of the data is undertaken thematically to ascertain pertinent features that doctors consider when making judgements about the adequacy of writing performance (relevant to research questions 1, 2 and 3). Furthermore, a quantitative analysis is performed to inform the basis of new standards and related cut scores. This is in response to research question four.
1.7  Outline of the thesis

This first chapter has highlighted the importance of communication in healthcare and the role of communication in applying for medical registration in Australia. The Occupational English Test (OET) which is the focus of this thesis, and in particular the OET Writing sub-test has been introduced. The motivation for the current study and related research questions have been outlined.

Chapter two is a literature review and provides further background to this thesis. First, the role of international medical graduates (IMGs) in workforce migration and globalization is introduced. Perceptions of IMGs in the workplace and IMG registration procedures and requirements in Australia are noted (including competent medical knowledge and satisfactory communication skills). Second, the importance of testing language readiness for healthcare in the Australian context is presented. The next section investigates the concept of language for specific purposes (LSP) testing including a theoretical justification for this approach and the debates which surround it. The place of the OET Writing sub-test within the context of LSP testing is then considered. The third section of the literature review investigates the topic of standard setting which is at the heart of the current study, beginning with a background to standard setting itself. Next, a categorization and evaluation of standard-setting methods is presented as this is crucial for choosing and justifying a method to be used in a study. Relevant studies and reports on standard setting are reviewed with an emphasis on language testing, the health professions (especially on tests used for assessing readiness to join the healthcare workforce) and the OET (limited studies). Lastly in this section, pertinent issues related to standard-setting research is investigated with a call for further qualitative research to counter these concerns. Since the current study can be seen as contributing to the validation of the OET, the final section of the literature review examines test validity including historical changes in the way it has been conceptualized leading into the current, argument-based approach. An account of an approach to validation that directs this study is provided with a particular focus on LSP testing and on standard setting in particular.
Chapter three outlines the methodology of this study with justification provided as to the choice of final standard-setting method that was utilized. Next, a report of pilot standard-setting workshops and a think-aloud protocol (TAP) trial is given together with an account of methodological refinements arising from this phase of the research. A description of the main study is offered including the participants, ethical clearance, instruments and procedures. Lastly in this section, a breakdown of the ‘decisions’ inference in an argument-based validity framework (employed in the current study) is presented to show the links between the relevant assumptions and specific research questions that were posed to the actual sources of evidence that were sought.

Chapter four firstly presents the qualitative results showing the basis for participants’ standard setting judgements and panellists’ final evaluation of the standard-setting procedure. Then, the quantitative results of the study are reported including the cut scores which emerged from the standard-setting exercise. These results are linked to the research questions posed at the outset of this study.

Chapter five discusses the study’s findings paying particular attention to the question, presented earlier in this chapter, of whether domain experts’ judgements can be considered as a valid basis for setting standards on the OET Writing sub-test. In answer to this overarching question, which is of relevance to LSP testing more generally, evidence related to the relevance of domain expert judgements to the OET construct, as currently operationalized, is reviewed. In addition, the question of whether the variability of domain experts’ judgements falls within acceptable limits is considered. Feedback from participants regarding their level of confidence in setting standards is also taken into consideration as to whether they are justified or not. The chapter concludes with a presentation of the new OET Writing sub-test cut scores, arguing that, in spite of some limitations, these constitute a defensible basis for determining the extent to which OET Writing sub-test candidates are equipped to undertake their professional role.
Chapter six presents the conclusion to the study. First, the study’s findings are summarized and framed around Knoch and Macqueen’s (in preparation) argument-based validity approach for an LSP test. Second, the study’s implications for the OET are introduced. Then, the overall contribution of the study to standard-setting research is stated. The study’s limitations are presented, and finally, future research areas are outlined.
Chapter 2: Literature Review

2.1 Introduction

The literature review firstly describes how workforce migration and globalization have affected doctor recruitment in Australia and how international medical graduates (IMGs) who commonly take the OET may be further relied upon to fill job shortages. Then, perceptions of IMGs in an Australian context are illustrated as these overall views may have an impact on domain experts’ interpretations of IMGs’ communication in the workplace. Also, IMG registration procedures and requirements in Australia are noted, and in particular, the approach taken to test language readiness for the healthcare setting in the Australian context.

The next section considers notions of communicative competence and specifically how these relate to language for specific purposes (LSP) testing, which is the conceptual basis for the OET, designed as it is to mirror the communicative demands of the healthcare workplace. A historical perspective of LSP testing and the place of the OET within this tradition are offered. A theoretical justification for LSP testing is outlined and the debates surrounding its utility as a basis for language test design and use are considered. Some thorny issues arising from the attempt to capture the communication demands of the real-world context are discussed and the central issue of domain expert involvement in LSP testing is examined.

The next section of the literature review considers standard setting. A categorization of a variety of standard-setting methods is given, as well as criteria for choosing a method for a standard-setting exercise, and an evaluation of several different possible methods for use in the current study. In addition, studies and reports on standard setting are presented with a particular emphasis on studies pertaining to language testing and the health professions. Also, recent standard-setting studies on the OET are reviewed. Potential issues in standard-setting research are identified (e.g. variability, method effects and domain expert involvement) and the importance of
qualitative research in standard-setting studies as a means of exploring the validity of participants’ judgements is highlighted.

Lastly, in order to attend to the validity concerns noted above, a section on test validity follows including an examination of historical changes in the conceptualization of validity. An argument-based approach to validity and the strengths/weaknesses of two key variants of this approach is investigated. How standard setting fits into an argument-based approach is then discussed. Finally, in this section, a validation method developed by Knoch and Macqueen (in preparation), with a specific emphasis on LSP testing that directs the current study, is provided.

2.2 International medical graduates (IMGs)

Firstly in this section, international medical graduates’ (IMG) place in workforce migration and globalization is noted; secondly, views of IMGs employed in an Australian setting are given; and finally, IMG registration prerequisites and procedures in Australia are stated.

2.2.1 Workforce migration and globalization

Workforce shortages in developed economies around the world frequently occur as a country’s population ages, birth rates decline and as labour markets shift and change. As a result of this, in Australia, there is a growing necessity for qualified doctors to be recruited from overseas countries. It has been argued that as Australia’s population continues to age and is faced with chronic illnesses, more and more international medical graduates (IMGs) will be asked to make up the healthcare worker shortfall (Australian Government Department of Health and Ageing, 2008, 2011; Barton et al., 2003; Hawthorne, 2012). As noted, approximately 40% of doctors with Australian registration were educated abroad (Department of Health and Ageing, 2011) and non-English-speaking background (NESB) IMGs from developing countries will likely rise in number (Mullan, 2005). It is probable that the Australian healthcare system will remain reliant on these health professionals (to some degree) into the future (Carver,
Thus, it is viewed as essential that IMGs are not only well-qualified, but also have satisfactory English language proficiency for them to be able to perform safely and effectually in the workplace.

The Australian Medical Association (AMA) has acknowledged the important contribution that IMGs have made and will continue to make to the Australian medical workforce, but regards shortages being filled by IMGs not as a long-term solution (AMA, 2004). More recently, some have also argued that the need for IMGs has become less urgent because of the growing number of medical students both trained in Australia and abroad. Birrell (2011) offers a view at odds with the idea of inadequate doctor supply and claims that as a consequence of steps to promote IMGs to work in Australia there are thousands who have difficulty in securing permanent jobs in the medical workforce. Borello (2016) reports that some leading doctors from Australia's medical community have pressed for an end to granting visas to IMGs as they claim the migration programs to address regional and remote skills shortages have been unsuccessful because IMGs might not be culturally ready to work in rural areas.

Nevertheless, Australia remains greatly dependent on IMGs to maintain adequate numbers of healthcare employees, and this is expected to continue because of their readiness to work in regions that habitually lack healthcare recruits, such as in remote and rural locations (Hawthorne, 2012; Short, Hawthorne, Sampford, Marcus, & Ransome, 2012). For instance, 46% of doctors in rural Queensland are IMGs (Hawthorne, 2012). Also, due to restrictions in the Health Insurance Act 1973, that came into effect in 1996-97, there is a requirement for IMGs wanting to work in private practice to be restricted to districts of workforce shortage (DWS) for up to ten years (Australian Government Department of Health, 2016). This will inevitably mean more IMGs moving to rural areas. It has also been noted that primary and acute health care provision will likely continue to be met by IMGs (Garling, 2008; Hawthorne, 2012).

In 2016, the Australian federal government health department made an official proposal (still under review) to change national immigration rules by requesting all medical occupations be taken off from what is called the ‘skilled occupation list’
(Parnell, 2016). In 2017, the federal Australian government assistant health minister announced plans for agencies that hire doctors for rural areas to be provided with financial enticements to employ Australian doctors over foreign-trained ones. This was to ensure “the doctor is the right fit for the community” as “official figures suggest Australia is headed for a doctor oversupply of 7000 by 2030, yet rural communities still suffer from a doctor drought” (“Move to swap foreign doctors for locals”, 2017). The situation in Australia is somewhat at odds with that in other jurisdictions such as the UK where Donnelly (2017) reports that the government there proposes to make it less difficult for IMGs and failing trainee doctors to become GPs due to extreme shortages and record vacancies – the UK government has vowed to increase GP numbers by 5,000 by 2020.

In the short and intermediate term, it seems likely that IMGs will remain a significant part of the Australian healthcare landscape, in spite of conflicting reports in the literature. This thesis draws on the identified need in Australia for well-qualified IMGs who have sufficient professional skills as well as language proficiency (the subject of this thesis) in order to be able to function effectively and safely in a variety of diverse workplaces (e.g., urban and rural settings). Also, IMGs both here and in other countries may be perceived as ‘problematic’ for a range of reasons; sometimes unfair bias, but also perhaps due to real linguistic and cultural difficulties they face in the workforce. These issues are examined in the next section.

### 2.2.2 Perceptions of IMGs in the Australian context

As noted above, international medical graduates (IMGs) are a permanent presence in the employment makeup of medical professionals currently working in Australia. Nevertheless, IMGs cannot be categorized homogeneously as they come from a variety of different circumstances and invariably have a range of medical experience from senior consultants to recently qualified graduates. Hawthorne (2012) notes the background of many IMGs in Australia is from countries including the United Kingdom, Ireland, New Zealand, India, South Africa, Malaysia, Singapore, China and the Philippines. Between 2005 and 2010 the UK represented around 30% of the
permanent migrant health professionals (HPs) who came to Australia and similar figures were seen for temporary migrants in the same period (UK, 27%) (Hawthorne, 2012). It can be observed then that many IMGs are from English-speaking backgrounds (ESBs) as well as significant numbers from non-English-speaking backgrounds (NESBs). It is important to differentiate between those who originate from countries that are similar to Australia in terms of language(s), background, culture and medical training and those from non-English-speaking backgrounds (NESB) (Pill, 2013) as the way they are perceived by various stakeholders may vary.

Pill (2013) argues that the term IMG often continues to be applied to overseas-trained doctors after they have satisfied all professional registration requirements because they may be regarded as still having a “deficiency (e.g., in language, professional knowledge or cultural competence)” and the “use of the term might therefore be contested” (p. 16). Bias against IMGs might occur in the workplace post-registration, as a study of complaints to medical boards in Australia (Western Australia and Victoria, prior to the 2010 nationalisation program) revealed that IMGs drew 24% more complaints and 41% more adverse findings than non-IMGs and also that complaints leading to adverse findings were not always an objective gauge of poor practice (Elkin et al., 2012). Further analysis showed that the probability differed significantly when comparing the country in which an IMG received his/her qualification because IMGs from Poland, Nigeria, the Philippines, Egypt, Russia, Pakistan and India had a considerably higher chance of attracting complaints as opposed to IMGs from 13 other countries who had no larger risk than Australian-trained doctors.

Another study that investigated perceptions of bias between IMGs and local Australian doctor graduates demonstrated no evidence of statistically meaningful variances when a questionnaire about attitudes was given to primarily Australian-born patients after a regular consultation (Harding et al., 2010). An additional study of bias conducted in Australia probed for evidence against overseas-born or overseas-trained doctors by asking future patients at a new local health clinic (Australian-born students of European heritage) to assess the competence and honesty of fictitious doctors stated to be submitting an application for work based on their CV and other related
information (Louis et al., 2010). The experimental study revealed that the medical education and workplace experience of doctors born and qualified in Pakistan were appraised less satisfactorily compared with Australian (Anglo-European) doctors. However, training in the UK was deemed as advantageous for both Pakistan-born and Australia-born doctors, which lessened the biased perception against the Pakistan-born group. These studies enquiring into IMG’s satisfactoriness in the workplace reinforce the view that they may be regarded as ‘problematic’ by locally-trained doctors and the wider community in which they operate. Therefore, the registration requirements they need to undertake and standards they need to meet are subject to scrutiny.

In the literature, a number of studies have focussed on some of the overall problems that IMGs might face in English-speaking healthcare situations (e.g., Baker & Robson, 2012; Dorgan et al., 2009; Hall et al., 2004). McGrath, Henderson and Holewa (2012) report that the initial period of employment in an English-speaking healthcare environment is a crucial time in relation to difficulties with clinical communication in practice and that IMGs need specific support to become used to an Australian healthcare context. Research has shown that language proficiency, communication skills, and Australian cultural understanding are just as vital as clinical ability to transition to successful healthcare practice (e.g., Chur-Hansen, 1997; Couser, 2007; Crawford & Candlin, 2013; Duncan & Gilbey, 2007; McDonnell & Usherwood, 2008; Pilotto et al., 2007; Saxena et al., 2006). Nevertheless, IMGs are not always characterized as being challenging or difficult to deal with. Srivastava (2008) observes the necessary sacrifice and perseverance of IMGs to achieve registration in Australia and Playford and Maley (2008) mention some of the benefits that IMGs give to medical contexts such as in teaching roles in rural and remote locations. However, there is little evidence of qualitative studies that specifically investigate the writing proficiency of IMGs in English and the effect on communicative competence in healthcare contexts.

Regarding perceptions of the OET from locally-qualified/experienced HPs in Australia, Macqueen, Yahalom, Kim and Knoch (2012) conducted a qualitative study into HP’s
knowledge and experience of OET test scores. It was demonstrated that there is “a level of trust placed in the testing authority” (p. 36) as the medical board representatives had not seen the OET test materials before the study’s interview. Wette (2011) also notes that the purpose and scope of the OET test may sometimes be misinterpreted by locally trained/experienced HPs as “employers can produce the false belief that higher scores on these tests will guarantee better clinical communication in the workplace even though such tests are not designed to directly measure the clinical dimensions of communicative competence” (Pill & Woodward-Kron, 2012, p.11). Macqueen et al. (2012) state that some of the doctors interviewed in their study considered that a passing score on the OET was “interpreted as indicating that an applicant is ready for the workplace in general” (p. 36) while others realized that newly registered IMGs would still need ongoing training in terms of context-specific language learning and that IMGs may still have language difficulties in the workplace. These studies indicate that there are diverse views about the purpose and consequences of test scores from the OET and its role in professional registration procedures for IMGs. IMGs both here in Australia and in other countries may be perceived as ‘problematic’ for a range of reasons; whether as a result of undue bias or because of real linguistic and cultural difficulties they may face in the workplace. The next section examines further pressure placed on IMGs regarding registration requirements and procedures in Australia.

2.2.3 Registration procedures and requirements

Greater recent examination of standards for Australian health professional registration by IMGs has been caused by several factors. As mentioned in the introduction to this thesis, the inquiry ‘Lost in the Labyrinth’ (House of Representatives Standing Committee on Health and Ageing, 2012) involved a review of the language tests utilized in Australia and the language proficiency standards of these tests. The evaluation of these standards may have been partly brought about by a perception that IMGs are still viewed as somewhat ‘challenging’ even after they have obtained full registration. IMGs are required to satisfy a range of prerequisites (including
communication skills) to professionally register and work which are primarily designed to ensure patient safety.

IMGs initially need to attain limited registration that allows them to carry out postgraduate training and/or supervised practice and also lets them be employed in designated areas of need (under supervision). There are three pathways for IMGs to receive limited registration. They must first decide their Australian Medical Council (AMC) (2017) pathway (i.e. choose to take the standard, competent authority or specialist pathway); have their credentials and qualifications certified; and pass an English language proficiency test at an acceptable standard (as noted, under Australian federal law, in order for IMGs to be registered, professional competence and English language proficiency are assessed separately (McNamara, 1996)). To acquire full registration with the Medical Board of Australia (MBA) (2014), the AMC Certificate is required (i.e. pass a medical multiple-choice test and clinical exam). Pill (2013) notes that further to the English-language test requirement, “the AMC examinations are delivered in English and therefore effectively constitute rather demanding language tests in themselves, but this assumption of a certain level of language proficiency in test takers is not recognized explicitly” (p. 18). The AMC Certificate involves completing a computer adaptive test (CAT) multiple choice questionnaire (MCQ) examination which includes “knowledge of the principles and practice of medicine in the fields of general practice, internal medicine, paediatrics, psychiatry, surgery, and obstetrics and gynaecology. It focuses on essential medical knowledge involving understanding of the disease process; clinical examination and diagnosis; and investigation, therapy and management”. In addition, attainment of an AMC certificate entails IMGs to pass a clinical examination that “assesses clinical skills in medicine, surgery, obstetrics, gynaecology, paediatrics and psychiatry. It also assesses ability to communicate with patients, their families and other health workers” (AMC, 2017). Furthermore, IMGs are required to obtain an appropriate visa, employment offer, and registration with the MBA.

As stated, one of the key criteria for the registration for IMGs in Australia is achieving the compulsory English language requirements (Australian Health Practitioner
Registration Agency (AHPRA), 2017). However, some English language skills test exemptions are permitted such as for those IMGs whose secondary and tertiary education were completed in countries where English is the first or native language (MBA, 2014). All other applicants must sit one of two English language skills tests – the IELTS, or the OET. As noted, the OET is currently recognized by registration authorities in Australia, New Zealand, Singapore, the UK, Ireland, Dubai and Namibia as satisfying the English language requirement (OET, 2017).

2.3 Testing language readiness for healthcare in the Australian context

While the IELTS and the OET are treated interchangeably by medical registration boards, they are very different in design. A comparison of the writing components of each test makes this clear. In the IELTS academic Writing test, candidates are required to complete two tasks: 1) a description, summary or explanation of a table, graph, diagram or chart; and 2) an essay in response to a problem, point of view or argument. In contrast, and as is the case with all LSP tests, the OET and its tasks are intended to reflect the actual communicative requirements situated in simulated real-world settings (Douglas, 2010). In the OET Writing sub-test for doctors (the test of inquiry in the present study), test candidates are presented with a set of medical case notes from which to write a referral letter or other professionally relevant text.

As stated earlier, despite the OET’s stated aim and design (based on the language demands of the healthcare workplace), it has been contended that the test is not entirely effective in accomplishing its ‘gate-keeper’ purpose (Lumley, 1995, 1998). Concerns have arisen that non-English-speaking background (NESB) IMGs may pass the test, yet continue to have difficulty with the required English-language proficiency to communicate safely and satisfactorily at work. As mentioned, low workplace communication standards are frequently credited to overly low passing standards on language tests like the OET and the suitability of these passing standards are the topic of continuing debate (e.g., Birrell & Schwartz, 2006). This debate is central to the current study and the question of suitable standards for the OET. Before addressing the specific question of what standards should apply, it is appropriate to consider the
broader domain of LSP testing, which underpins the design of the OET, including the claims that LSP tests make about measuring language use for specific occupational purposes and the debates which surround their design and use.

2.4 Language for specific purposes (LSP) testing

This section presents an overview of the field of language for specific purposes (LSP) testing. First, the notion of communicative competence and how it has informed thinking about LSP assessment is introduced. Second, theoretical arguments for LSP tests and their practice are reviewed. Then, a range of opinions from the language testing community concerning the validity of LSP tests and their usage are examined. Some thorny issues that LSP testers have grappled with are discussed, and particularly the notion of LSP test authenticity or the fidelity of the test to what matters in the relevant domain of language use. The role of domain experts in achieving such authenticity is contemplated, and the importance of discovering what domain experts’ value when making decisions about standards on a LSP test is emphasized. Finally, the place of the OET within the LSP testing tradition is considered.

2.4.1 Communicative competence

Views of communicative competence affects how the construct of ability that a language test sets out to measure is conceptualized, how the test is designed and how test results are interpreted. These views vary widely, reflecting different positions on what it truly means to know a language and be communicatively competent in it. Debate continues in the literature as to what features are actually required to know a language and be proficient in it because how these aspects of language knowledge and communicative competence interrelate may be indistinct (Chalhoub-Deville, 2003). Research in second language acquisition, evolutionary biology and neuroscience (Hulstijn, 2011) has furthered understanding in this sphere, but how language is acquired/learned or how language knowledge is operationalised in performance is still not precisely understood. A number of terms have been developed in the area’s conceptualisation such as ‘communicative competence’
‘language proficiency’ and ‘communicative language ability’ (see e.g., Bachman, 1990; Bachman & Palmer, 1996, 2010; McNamara, 1996; Purpura, 2008) and is discussed further below. A broad question that needs to be considered is whether an LSP test adequately captures the construct of communicative competence as it plays out in a real-world setting. The test construct will be reflected in the task design, and also in the rating criteria and the way they are interpreted and operationalized by raters. It may also, as seen later, be reflected in the features that domain experts attend to when setting performance standards on the test.

Notions of communicative competence have been an arena of interest in language acquisition and language testing for many years. From early developments in the study of first language acquisition, a number of theories have been offered relating to the term ‘communicative competence’. Hymes (1972) first proposed the idea of communicative competence in response to Chomsky’s (1965) initial formulation of the notion of ‘linguistic competence’ in his model of ‘Generative Grammar’. Chomsky’s (1965) theory set out to explain the language knowledge possessed by an ‘ideal’ first language user and centered not on a first language user’s real-life language use that might incorporate errors of performance, but rather on an ‘idealized’ language system. Chomsky (1965) noted the distinction between competence and performance when explaining the degree of an individual’s proficiency in their skills or knowledge. He argued that competence is the underlying level of skills or knowledge that a person might have. This competence could be demonstrated through performance in either an enhanced or hindered way, depending on cognitive and affective factors (e.g., excessive confidence or memory issues) that may have an effect on an individual at a particular moment of performance.

Hymes (1972) argued for a different conceptualisation and considered Chomsky’s linguistic competence theory to be inadequate due to its lack of attention to social aspects and communicative conduct. Whereas Chomsky’s interpretation of competence focussed on an individual’s demonstration of knowledge and whether a sentence was ‘formally possible’ or not, Hymes also included knowledge of whether a sentence was ‘feasible’ and ‘appropriate’ to be used in a specific context, and if it was
‘actually used’ or not. Hymes’s notion takes account of both a learner’s implicit or ‘tacit’ language knowledge and also the facility to use it successfully, which he termed ‘ability for use’ (Hymes, 1972, p. 282). Ability for use incorporated cognitive and non-cognitive components as they can both affect performance in real-world settings. Ability for use, rather than underlying competence, is regarded as a gauge of language proficiency; however, it is also associated with social and cultural contexts (Douglas, 2010).

The concept of communicative competence put forward by Hymes (1972) was further extended into second language acquisition teaching and testing in a ‘communicative approach’ by researchers such as Widdowson (1979), Canale and Swain (1980), Savignon (1983) and Canale (1983). Canale and Swain’s (1980) purpose was to establish a theoretical basis that incorporates “a clear statement of the content and boundaries of communicative competence” (p. 1). Their framework takes into account three factors involved in communicative competence: grammatical, sociolinguistic and strategic competence and places psychological features as universal aspects that are not specific to language. This was to some extent different to Hymes’s theory because this construct discounts the idea of ‘ability for use’ and regards psychological features as universal aspects (i.e. characteristics not specific to language) as they are not relevant and do not require ‘conscious learning’ in a natural performance (Canale & Swain, 1980, p. 16). This construct was further extended by Canale (1983) who regarded world (content) knowledge as distinct from the concept of communicative competence because it interrelates in undetermined ways with further organizations of skills and knowledge (e.g., world knowledge) and also with a theory of human action (e.g., taking into account aspects like personality and choice). Also included by Canale (1983) in his model of communicative competence was discourse competence, including the sub-components of cohesion and coherence.

Another adaption of Hymes’s (1972) and Canale and Swain’s (1980) theory of communicative competence was proposed by Bachman (1990) and Bachman and Palmer (1996, 2010) in the form of a model of ‘(communicative) language ability’. Their psycholinguistic or cognitive model (based on empirical research) has been
updated several times since the 1980s and is directly related to language testing. Language ability in their view encapsulates two key constituents: ‘language knowledge’ and ‘strategic competence’ (Bachman & Palmer, 2010). Language knowledge, in Bachman and Palmer’s model, encompasses organisational knowledge (grammatical and textual) and pragmatic knowledge (sociolinguistic and functional). Strategic competence includes the metacognitive strategies of goal setting, planning and appraising. Also, in this model, additional characteristics may interrelate with language ability such as the personality of the test candidate and features of the assessment task and context. This theory of communicative language ability includes the psychological features excluded by Canale and Swain’s (1980) framework because Bachman and Palmer (2010) see that these strategic devices perform a necessary function. However, both Canale and Swain (1980) and Bachman and Palmer (1996, 2010) regard background or content knowledge as an uncontrolled variable and hence remove it from their construct definitions. In contrast, Douglas (2000), when defining the construct foundation of an LSP test, characterizes content knowledge as a vital feature: “the interaction between language knowledge and content... is perhaps the clearest defining feature of LSP testing” and content or “background knowledge is a necessary, integral part of the concept of specific purpose language ability” (Douglas, 2000, p. 2).

Kramsch (1986) looks at communicative competence from another approach in a sociolinguistic position formed from observations of second language learning and teaching, which is somewhat in contrast to Bachman and Palmer’s (2010) cognitive perspective. Her notion of ‘interactional competence’ takes a social view of interaction and focusses on the various ways that language shapes reality. Language from this stance is regarded as a “dynamic process of communication” dependent on context where meaning is constantly negotiated between participants instead of a “static content structure” in which interactional skills required to cope with intercultural differences are only attained when learners have reached a higher proficiency level (Kramsch, 1986, p. 368) or as termed by Jacoby and Ochs (1995) in a ‘co-construction’. From Kramsch’s perspective, language learners need to be aware of their own language norms and have a capacity to create some distance from these
conventions which permits them to predict possible occurrences of miscommunication. Jacoby and Ochs (1995) argue that interactional competence is regarded as more related to performance than underlying competence because “it is communicative competence, ... but as it plays out in all its incredible complexity as people go about managing their identities, their relationships, and their lives” (p. 179). Chalhoub-Deville (2003) likewise regards context similarly to Kramsch (1986) as a vital component of the construct instead of a concern to be studied independently. The next section examines theoretical arguments for LSP tests and their usage.

2.4.2 Theoretical justification for LSP testing

LSP tests started to be developed in the 1970s as Hymes’ (1972) notion of communicative competence became more prevalent in second language teaching and learning contexts and as performance assessments in work environments became more common (O’Loughlin, 2008). Douglas (2000) notes the ‘Temporary Registration Assessment Board’ (TRAB) (later revised as the PLAB – Professional and Linguistic Assessment Board) as one of the first situations of a language test being used for a specific purpose. It was established in 1975 by the General Medical Council (UK) as a test for IMGs seeking to gain temporary medical registration in Britain. The assessment utilized specific medical related language tasks based on a needs analysis as well as a test of professional skills and knowledge. The test was developed by language specialists and importantly with input from medical experts. Rea-Dickins (1987) argued that this type of partnership in the TRAB’s development “would seem to be a prerequisite for the design of a ‘special purposes’ test as the domains incorporated within the specialist area go beyond those in which the linguist – independently – is competent to make judgements” (p. 196). The test developers also sought to include test tasks that combined language proficiency as well as authentic and appropriate background/content knowledge (e.g., the writing task involved patient case notes and a letter of referral) (Douglas, 2000).

Douglas (2000) contends that by definition LSP tests are communicative and “it is impossible to distinguish LSP testing from communicative language testing”
(Sajavaara, 1992, p. 123). Therefore, from this point of view, LSP tests are in effect magnified communicative language tests not only because they include different features of communicative competence, but also as the entire undertaking is concerned with testing peoples’ ability for use (from Hymes’s perspective) in a particular real-world setting. LSP tests contain tasks that are intended to draw out communicative and interactive performances to replicate real-life situations (Fulcher, 2000). Hence, LSP assessments centre on how well test takers can demonstrate their communicative capability in a certain domain.

The basis of LSP testing, as articulated by Douglas (2000), centres on the notion that unique types of a language are able to be grouped together within definable ‘discourse domains’. When considering the context of communicative language use, Douglas (2000) proposed it as a psychological construct that is affected by linguistic and physical contextualization prompts. According to this viewpoint, what is important is how a language user understands the contextualization prompts in attendance in a communicative event rather than the external context in isolation. This suggests an “internal view of context as a construct created by language users for the interpretation and production of language” (Douglas, 2004, p. 25). This internal interpretation of context is termed ‘discourse domain’ by Douglas and Selinker (1985a, 1985b) who define the concept as “a cognitive construct created by a language learner as a context for interlanguage development and use” (Douglas, 2004, p. 25). Selinker and Douglas (1985) when focusing on second language acquisition describe a discourse domain as “internally-created contexts, within which...IL (interlanguage) structures are created differentially” (p. 190) and when concerning language testing “each test taker creates for him or herself an internal context within which he renders the text intelligible” (Douglas & Selinker, 1985a, p. 206).

As language performances differ depending on the situation and the discourse domain, LSP tests are meant to simulate the authentic communicative requirements of test takers in that context (Douglas, 2010). For example, in fields such as medicine and nursing, assessment tasks are intended to meet a functional need for testing in that specific domain (Davies, 2001). From a simulated performance situation, the goal
of an LSP test is to predict a test candidate’s language proficiency in a real-world setting. LSP tests frequently comprise domain-specific language and performance-based tasks (i.e. direct tests that hope to mirror situations from real workplaces) (Douglas, 2010). This is different from general proficiency testing where a less direct approach might be taken, given that the test is designed to cater for a range of contexts. Douglas (2010) notes that since LSP tests aim to mirror the real communicative requirements of individuals in that circumstance, in the preliminary stages of test development a target language use (TLU) study is often undertaken. Understanding language use in the context of concern serves to inform test design to make it more likely that test tasks or items will elicit relevant language skills. LSP tests, should then in theory, estimate the way in which a test taker will perform in the target domain with greater accuracy than general English language tests (Douglas, 2010; Elder, 2001).

LSP tests are often used for high-stakes purposes such as to determine readiness to enter the workforce such as in the health professions, where effective communication skills are critical for patient safety (Basturkmen & Elder, 2004). In other high-stakes fields such as aviation (again for public safety reasons) there is a requirement by the International Civil Aviation Organization (ICAO) for pilots and air traffic controllers to have an ‘operational’ standard of English-language proficiency and several tests have been recognized by national aviation authorities to assess English-language competence (e.g., RMIT English Language Test for Aviation (RELTa)) (International Civil Aviation Organization, 2017). Because the stakes are high for tests like these, any failure to accurately replicate the work setting might produce damaging ‘washback’ where test candidates could be given a conflicting awareness about their future work environment (Basturkmen & Elder, 2004). Such failure would also jeopardize the validity of test scores, which could not be taken as accurate representations of readiness to work in the target domain. Therefore, LSP tests should endeavor to replicate the real-world context as reliably as possible (see below in regard to test authenticity).
There are a number of debates surrounding LSP tests and their usage. These discussions have continued since LSP testing’s initial foundation and are examined in the next section.

### 2.4.3 Debates about LSP testing

A range of sometimes conflicting views regarding the merit of LSP tests and their use have been put forward by the language testing community. Discussion of these concerns and debates are organized around the three issues of specificity, separability and authenticity identified by Douglas (2001b) as particularly challenging for the theory and practice of LSP. Specificity refers to the problem of identifying what is specific to the target language use domain (i.e. a problem of domain definition and distinction between domains). Inseparability concerns the difficulty (on tests designed to measure language proficiency in specific domains of use) of separating language from discipline-specific knowledge. Authenticity involves the quest for ‘life-likeness’ in what is inevitably the artificial environment of a test.

#### 2.4.3.1 Specificity

Language tests for specific purposes base their assessment content and method on analyses of language use in the specific target language domain (Bachman & Palmer, 1996; Douglas, 2000, 2001b; Stansfield & Wu, 2001) that the test is designed to capture. LSP tests have therefore been claimed to be more accurate predictors of real world performance because they are more specific in their orientation and hence more contextually sensitive than their general proficiency counterparts (Douglas, 2010; Elder, 2001). Despite this claim, there is at present little empirical evidence to point to this (Basturkmen & Elder 2004; Davies, 2001; Douglas & Selinker, 1992, 1994; Read & Wette, 2009). Davies (2001) concludes that even though LSP testing may still be “worth pursuing”, for face validity reasons, it “remains of uncertain value and, indeed, has not proved itself to be more valid than a general proficiency test” (p. 144). Conversely, Douglas (2000) contends that there is also a lack of evidence that a general
language proficiency test is an adequate determiner for LSP contexts such as healthcare.

Douglas’s (2000) scepticism about general proficiency testing is based on the belief that ability estimates derived from one performance are not necessarily generalizable to those derived from performance in another context, as according to the discourse domain theory discussed above, a candidate’s language performance will change with situational context and type. Because of this, he advocates that test candidates should be provided with specific and authentic assessment tasks where subject-area knowledge and language proficiency interrelate with the assessment content matter. In sum, he argues that with test tasks that typify the target context, a better case can be made for valid measurement of a candidate’s ability.

Davies (2001), however, holds the view the LSP assessment has a "lack both of discreteness and of coherence" (p. 137) arguing that clearly delimited domains of language use according to context may be difficult to distinguish. Accordingly, Davies (2001) claims that testing LSP proficiency could be seen as merely evaluating general language proficiency with specific technical terminology incorporated. Similar concerns are expressed by Elder (2001) when referring specifically to the testing of teacher proficiency. She demonstrates the problem of establishing boundaries between the language requirements of different teaching contexts (e.g., primary school teaching and university teaching) which may differ widely from one another in their language demands.

Douglas (2001b) concedes that there are challenges associated with specificity of language use and questions whether it is possible to satisfactorily encapsulate the specificities of particular domains in a language test. He argues that in spite of the thoroughness of a target language use (TLU) needs analysis, the associated test “might not be truly representative of the target situation, since there are simply too many possible variations of situation to cover adequately in a test situation” (p. 49). Thus, some test candidates may experience distress if the assessment content or tasks do not seem to match the specific circumstances and demands of their own professional
context. Issues might also occur if assessments designed for a specific application by one regulatory authority are then used by another (e.g., a test containing content or informal terminology specific to Australia which is later adopted for use in Singapore). This could be a possible concern for test developers as LSP tests are employed in more jurisdictions around the world.

In summary, while a case can be made for specific purpose tests, there are uncertainties surrounding the discreteness of different domains of use. There are also practical challenges for test developers in achieving the right degree of specificity given the variability of language use within a particular context.

### 2.4.3.2 Inseparability

Regardless of the purpose of a language assessment (whether general or specific), language is the means and the aim of the test. An unresolved concern identified by Davies (1995) is that there needs to be content in a language test because “knowledge needs language to encode it” and at the same time “language needs knowledge or content to give it meaning” (p. 11). However, as indicated by Davies (1995), in LSP testing, the situation is rendered more complex for its specific-purpose emphasis necessitates that content knowledge is part of what is being assessed and not just a vehicle for eliciting information about language ability. This is at the heart of another issue raised by Douglas (2001b) – that of the inseparability of language from other knowledge (e.g., doctors’ clinical knowledge) and skills utilised by test candidates in performance on LSP test tasks. Language ability and content knowledge are, according to Jacoby & McNamara (1999), “inextricably intertwined” (p. 234). Douglas (2000, 2005) likewise claims that assessing language proficiency and content knowledge in tandem is a fundamental and indispensable feature of an LSP test construct. This is at odds with the view of Messick (1989), who argues that when developing a general language proficiency test, the test designers aim to avoid taking non-related content knowledge into account as this should be seen as irrelevant to the construct. Davies (2001) also contends that “LSP testing cannot be about testing for subject specific knowledge”, however it should rather be “about testing the ability
to manipulate language functions appropriately in a wide variety of ways” (p. 143). In LSP testing the problem still exists of how and indeed whether to separate language ability from subject-specific knowledge. There is continued debate as to whether separation is adversative to LSP testing or an imperative.

These different views are encapsulated in McNa mara’s (1996) distinction between ‘strong’ and ‘weak’ performance tests. In a ‘weak’ performance test the assessment criteria relate to the test taker’s underlying language proficiency usually in the context of a simulated real-world task. In such assessments, the task is only a medium for obtaining a language sample (McNamara, 1996). A weak test is more in line with Messick’s (1989) and Davies’s (2001) view that the test task, however realistic, is only used as a ‘vehicle’ for assessing language. On the other hand, the assessment criteria in a ‘strong’ performance assessment imitate those used in an actual work situation and thus do not rely on language proficiency alone, yet may also encompass other aspects of professional communication. In strong performance assessments, the task is not only the medium, but also the object of measurement (i.e. the task is being judged on how well it has been achieved from a professional viewpoint) with aspects apart from language emerging as relevant features. A strong test is more in line with Douglas’s (2000, 2010) view, in LSP testing, that the task performance itself becomes the focus of assessment and any aspect of communication (including content knowledge that is relevant to the success of that performance) is taken into consideration in judging the test candidate’s proficiency. For example, studies in the evaluation of aviation English in which qualified pilots consider technical knowledge when assessing the communicative ability of their peers (e.g., Knoch, 2014).

LSP tests may sit uneasily on a cline between strong and weak performance assessments leading Elder (2001) to point to “the indeterminacy of LSP, as currently conceptualized, as an approach to test development” (p. 149). Pill (2013) acknowledges this indeterminacy, but argues that a weak LSP test or “a language test with ‘traditional’ language assessment criteria may be ‘strengthened’, that is, moved along a cline from weak towards strong, in McNamara’s (1996) terms, by including for assessment further aspects of performance that are meaningful in the workplace.
2.4.3.3 Authenticity

Authenticity is regarded as a key component of communicative language testing and in particular for the development and validation of LSP tests (Bachman & Palmer, 1996). Communicative assessment tasks that are related to test candidates’ experience outside the test environment are claimed to be more authentic than tasks that are dissimilar to real-life contexts (Ross & Berwick, 1992). This means the nearer the criterion measure is to the assessment context, the greater the authenticity of the test and the more effectual predictor of future performance it is liable to be. However, it must be taken into consideration that authenticity is not an absolute quality and is a question of degree as all assessments are to some extent inauthentic because of the inevitable artificiality of the assessment situation.

Widdowson (1979) divided authenticity into ‘genuine’ instances of discourse “designed to meet a communicative purpose, directed at people playing their roles in a normal social context” and situations designed as “a contrivance for teaching language” (p. 89) and also for language testing. He argued that it is important when evaluating the authenticity of a test task to consider the type and form that test input information should take and how it is utilized according to the test developer’s aim. Bachman and Palmer (1996) extended Widdowson’s concept of authenticity for test purposes by considering both ‘situational authenticity’ and ‘interactional authenticity’. The former pertains to task development and the extent to which the features of the test task and the conditions under which it is performed resemble significant aspects of the target language use (TLU) context as established through needs analysis. The latter term integrated Widdowson’s (1979) notion of ability for use (also noted by Hymes, 1972) and went beyond the features of test design by incorporating the way in which test takers engaged with assessment tasks and the degree to which performance on the assessment in question generated the skills
concerned with communication in the real world. Achieving interactional authenticity, according to Douglas (2000), is a key condition for defensible decision-making.

However, given the artificiality of the testing environment mentioned above, the task of characterizing or capturing relevant features of the LSP domain on the test (situational authenticity) and of engaging the kinds of skills that are deployed in real-world domains (interactional authenticity) is a challenging one. Furthermore, as the realities and limitations of test design will unavoidably influence authenticity, the question of how skills are deployed in performance should be examined in test validation research (Lumley & Brown, 1996; Spence-Brown, 2001).

Another aspect of this authenticity challenge is that language experts who design and implement LSP tests may not have access to what is important for successful performance in the particular domain. Hence, it may be necessary to look more closely at how those who work (domain experts) in the relevant occupational area conduct themselves in their workplace context and also how they conceptualize language proficiency. A key question concerns the problem of how to define what matters for successful performance in an LSP test and whether this aligns with the views of subject-matter experts. Recognition of the value of canvassing the perspectives of insiders (domain experts) in the specific domain has led to theorizing around the notion of ‘indigenous assessment’ which is further examined in the next section.

2.4.4 Domain expert involvement in LSP research

The idea of understanding what matters to domain experts and their ‘indigenous assessment’ procedures is an area of relatively recent enquiry in LSP testing research. The distinct conceptualizations of communicative competence characterized earlier reveal the size of the challenge for language test designers in encapsulating all the areas of knowledge and ability that need to be tested. The complexity of the previously mentioned frameworks of communicative competence have also highlighted the limits of what is testable. Chalhoub-Deville (2003) proposes that theorists are now searching for explanations that are suitable to a specific context
rather than comprehensive theories. Her reference to McNamara’s (1997a) proposal that “close analysis of naturally occurring discourse and social interaction may reveal the standards that apply in reality in particular settings, which may not be at all those proposed by language testers” (p. 457) is pertinent to the purpose of the present study.

A number of studies have been concerned with the criteria employed by domain experts in evaluating communicative competence (see e.g., Brown 1995, Elder, 1993, Douglas & Myers, 2000; Jacoby, 1998). Jacoby (1998) investigated the notion of ‘indigenous assessment practices’ through ethnography and conversation analysis of a group of academic physicists when they were providing peer feedback to each other in preparation for conference presentations. In Jacoby’s PhD study, the group’s members were from both English-speaking backgrounds (ESB) and non-English-speaking backgrounds (NESB), and although they gave similar feedback on language matters, other issues more relevant to the specific context were the focus of concern. The findings showed that non-language experts (physicists) paid far more attention to non-linguistic features rather than linguistic errors. New members to the group adhered to the norms and practices of that specific professional activity and collectively “call upon their own indigenous members’ methods of practical reasoning and on a rich inventory of tacitly known assessment criteria” (Jacoby, 1998, p. 311). Studies such as this reveal that “assessment criteria used by domain experts in judging actual communication are considerably different from the conventional linguistically-oriented criteria” used in most language tests (Elder et al., 2017, p. 3). Many of these studies have shown that non-language specialists have a tendency to evaluate the communicative competency of L2 speakers somewhat differently and are less concerned with linguistic accuracy whereas language experts are more typically sensitive to linguistic form and are severer on errors.

Jacoby and McNamara (1999) further explored this issue and state that to have a thorough understanding of what is involved in domain specific performance we need to gain proximity to the domain and only then can we account for what matters to the people concerned. This is because specific-purpose performance “is by definition task-
related, context-related, specific and local” (p. 234). Jacoby and McNamara (1999) refer to the idea of ‘indigenous assessment criteria’, or the values that underlie everyday judgements of performance by domain experts in actual workplace situations. For example, health professional experts may have awareness, whether conscious or unconscious, of the relevant indigenous assessment criteria through their built-up workplace experience and knowledge. Furthermore, Jacoby and McNamara (1999) claim that “studies of naturally occurring ‘indigenous’ socialization and assessment practices in professional settings, can provide more direct access to what counts as communicative competence in particular contexts” (p. 214). This has led to a focus on criteria that underpin domain experts’ judgments of performance and the importance of capturing these within the test construct (Jacoby & McNamara, 1999).

LSP assessment research informed by Jacoby (1998) and Jacoby and McNamara’s (1999) notion of indigenous assessment has thus far focused either explicitly or implicitly on the criteria for judging test performance. Fulcher, Davidson and Kemp (2011) employed what they term an ‘indigenous perspective’ to synthesize a range of earlier “discourse, focus group and narrative studies” (p. 7) exploring the particular travel service and marketing domains of interest. From their review, they claim to have acquired a thorough understanding of the assessment context and discovered what domain experts value. Based on this understanding they devised a construct of interactional competence, comprised of the components of discourse competence, competence in discourse management and pragmatic competence which were proposed as criteria for assessing performance on tests of communicative ability for those working in the travel industry or marketing services. Pill (2013, 2016), also working in the indigenous assessment tradition, undertook a close analysis of medical professional’s commentary of the strengths and weaknesses of a set of training videos showing interactions between patients and trainee doctors. From this analysis he derived a conceptual model of what health professionals value in clinical consultation which led to the formation of new professionally relevant assessment criteria for the OET Speaking sub-test.
The importance of recognizing what is valuable to subject-matter experts when making judgements and drawing on their own indigenous assessment criteria has been highlighted in this section. However, whether the values that emerge in an artificial research environment are indeed those that underpin their attitudes and behaviours in the real-world setting is questionable. McNamara’s (1997a) view of indigenous assessment is that it must be captured ‘directly’ in the workplace context to be considered as authentic, yet this stance presents practical and methodological challenges. For example, Elder and McNamara’s (2016) paper (which utilized data drawn from a larger, previously mentioned ARC-funded project on the OET Speaking test) investigated different methods to elicit domain experts’ (physiotherapy educators and supervisors) viewpoints on what mattered for successful communication in a work context. The study compared qualitative data from three situations – two were in the actual workplace in which communication skills are routinely judged and feedback is given instantly and more naturally rather than being elicited. The other situation was a simulated workplace environment that involved a video-recorded patient/therapist roleplay. The recorded roleplays were then shown to domain experts and they were asked to review the communicative strengths and weaknesses of each spoken performance. A qualitative analysis revealed that the character of subject-matter expert feedback varied considerably between each setting. The routine feedback that was given by supervisors to trainees in actual work contexts offered little insight into communication skills (e.g., due to issues such as constrained access, a lack of privacy and time pressures). In contrast, even though the research-oriented workshop context was less authentic, it generated a much richer understanding of physiotherapists’ assessment of trainees’ communication in a workplace-like situation, encouraging a “meta-awareness of criteria” that was not as forthcoming in the real workplace environment (Elder & McNamara, 2016, p. 169). Therefore, given the above-mentioned difficulty of accessing what domain experts’ value in real workplaces, it may be more fruitful to utilize other methods of qualitative data elicitation outside the work setting such as focus group workshops or verbal reports as adopted in the current study.
The above discussion of authenticity in LSP testing has acknowledged the importance of domain expert involvement at all stages of test design, beginning with needs analysis and extending to the assessment criteria which are used to judge performance, where the notion of indigenous assessment has been influential. However, there has been relatively little focus on domain expert involvement in the final standard-setting stage of test development where decisions are made about how much ability is enough for satisfactory performance in the context of concern. Standard setting also requires that domain experts make value judgements, but the criteria they employ in making decisions about performance levels are often unclear. Elder et al. (2017) argue that “incorporating ultimate arbiters’ perspectives into test development enhances the validity of language-for-specific purposes tests” (p. 3). Therefore, for reasons of test validity, further investigation and particularly qualitative research into the values that underlie domain experts’ judgements when making standard-setting decisions on high-stakes LSP tests is warranted and is explored in this study, using procedures that are not dissimilar to those adopted in previous research within the indigenous assessment paradigm.

The following section examines some of the LSP assessment issues raised above as they apply specifically to the OET. Some relevant OET-focused research studies are reviewed, with a particular emphasis to those that centre on the writing component, which is the focus of the current research.

2.4.5 LSP testing and the OET

The Occupational English Test (OET) has been regarded as an exemplar in the LSP testing field (Douglas, 2000) and numerous checks and balances have been undertaken (see below) to justify its creation and continued usage. The OET encompasses all the aspirations of LSP testing, but also some of the complexities and areas of tension that have been discussed previously. Even though there has been considerable research conducted into various aspects of the OET, some areas have been under-investigated and hence warrant further inquiry.
In developing the OET, McNamara (1990, 1996) conducted an initial needs analysis to establish the target language and content demands of the health professions under consideration. According to Davies (1984), in language testing, the content, layout, design and types of test tasks chosen are critical in the development of these types of LSP tests. McNamara (1996) used a variety of means to inform the design process such as ‘interviews’ with educational trainers of local and overseas-trained health professionals, a ‘questionnaire’ for persons with direct health workplace experience, ‘direct observation’ of healthcare workplaces and an ‘analysis’ of crucial communication tasks in healthcare settings. The needs analysis noted the following ‘real-world’ writing responsibilities in healthcare settings as the foundation for the OET Writing sub-test: a) recording items in patient hospital records; and b) communication with other health professionals (McNamara, 1990). While there are no publicly available test specification documents from Cambridge Boxhill Language Assessment due to confidentiality and security reasons, broad test specifications for the Writing sub-test formulated by McNamara (1990) were:

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

McNamara’s (1990) initial needs analysis for the OET attempted to ground the assessment content and language use in the specific language domain, as is the case in all LSP tests. However, the problem exists that while McNamara’s needs analysis aimed to achieve appropriate specificity of content, in actuality this was difficult to realize given the varying demands of the different health professions (e.g., medicine, nursing and physiotherapy). This concern is linked to the specificity issue mentioned previously by Bachman and Palmer (1996), Douglas (2000, 2001b) and Stansfield and Wu (2001). An example of this specificity problem is the OET Writing task which may be pertinent for doctors (i.e. in this instance a letter for referral), but which has dubious relevance to the nursing profession.
To address any authenticity concerns about the OET Writing task, in his preliminary needs analysis, McNamara (1997b) included the views and scrutiny from healthcare industry professionals. These domain expert opinions informed that in a number of clinical settings, referral and reply letters are an essential part of written communication for information exchange between health professionals (Berendsen, Kuiken, Benneker, Meyboom-de Jong, Voorn & Schuling, 2009; McConnell et al., 1999; Piterman & Koritsas, 2005). More recently, as part of a study that investigated current writing demands in healthcare settings, Macqueen et al., (2012) canvassed key stakeholders’ perceptions of the OET. The qualitative study focussed on senior doctors and nurses who manage EAL (English as an Additional Language) staff and professional medical and nursing/midwifery board representatives of their opinion of the OET test task itself and of its usefulness in the workplace. The stakeholders were from hospitals and aged-care residences in rural, regional and metropolitan areas of Australia. For doctors, the study found that “the task of composing a formal handover letter, e.g., referral or discharge, from case notes was highly relevant” and “was closely linked to their daily practices” (Macqueen et al., 2012, p. 18). Medical experts thought that the task of “selecting, synthesising and transferring information into formal prose for receipt by another health professional to be worthwhile, although the real-life task is far more complex” (Macqueen et al., 2012, p. 18). However, as mentioned, some health professionals, for example nurses, do not write referral letters as part of their everyday work routine. Therefore, for this profession at least, neither the task and the assessment criteria, are truly authentic and mirror real-life work procedures.

In regard to the development of the criteria for assessing the OET’s productive skills, these were initially founded on a test for United States foreign affairs officers (Foreign Service Institute (FSI) Oral Interview) that considered language proficiency compared to the language use of native English speakers (Douglas, 2001a; Jacoby & McNamara, 1999; McNamara, 1990; McNamara, 1996). The OET assessment criteria that emanated from the FSI were modified to have a more “communicative orientation” (Jacoby & McNamara, 1999, p. 218). However, as they do not explicitly evaluate the communicative practices vital to the healthcare context, this may provoke doubts concerning test construct validity. In the case of the OET, health professionals’ criteria
should be elicited and applied because if language experts’ criteria are solely relied upon, the test construct may end up being underrepresented or misrepresented (Pill, 2013). Douglas (2001a) and Hamp-Lyons and Lumley (2001) argue that the development of professionally relevant criteria is still one of the most pressing issues for research and practice in LSP.

According to Jacoby and McNamara (1999), the OET’s assessment criteria have a communicative focus, but due to an Australian federal government legislative requirement, they are confined to facets of language performance and no other aspects of healthcare communication that may be considered important by stakeholders in real-world contexts. As mentioned, there has been a recent ARC-funded project related to the OET Speaking sub-test (Elder et al., 2013) (along with the current ARC project concerned with the OET Writing sub-test) which have both explored health professionals’ indigenous criteria and pushed the boundaries of research in this area. Both studies have yielded a new set of professionally relevant criteria for the OET, which will in due course be used to assess performance.

In the final stages of developing the OET Writing sub-test materials, McNamara (1990) conducted a trial to analyse the new task’s writing criteria and rating grid, to assess inter-rater reliability and to set passing scores. The trials were conducted with overseas-trained medical and nursing practitioners and two ESL teachers as raters who used the new rating grid and worked independently. Another trained rater used the earlier OET version rating scale from the Australian Second Language Proficiency Ratings (ASLPR); a tool widely recognized in Australia at the time. A constraint on the study was that the Council on Overseas Professional Qualifications (COPQ) recommended that a passing standard and cut score be set at an equivalent of ASLPR level 3, which was retained. On the new six-point rating scale the passing cut score was set arbitrarily in the range of 6.75 to 9, without subject-matter expert involvement. In the trial, McNamara (1990) found that inter-rater reliability was limited, perhaps due to the ESL teacher raters not having the required clinical knowledge and background experience and advocated that future training of raters by health professionals themselves should be a priority. In addition, the use of the
ASLPR scale was not deemed suitable as the criteria on the general proficiency scale were in conflict with the new OET rating scale. It was also recommended that in setting passing standards, input from domain experts from the health professions should be part of the process. This had not been the case in the trials, hence jeopardizing the authenticity and relevance of these standards in relation to the demands of the healthcare domain – an issue which this study sets out to remedy.

In summary, since the introduction of language for specific purposes (LSP) testing a number of noted different viewpoints to describe its conceptualization have been considered. This has led to a range of debates in the language assessment community about the theoretical basis for LSP testing. These have centred mainly on what Douglas (2001b) sees as central areas of tension for LSP theory; namely the notions of test specificity, inseparability and authenticity. Of particular relevance to test authenticity is the idea of domain expert involvement, not only in the early needs analysis stage of test task design, but also in subsequent phases, including the development of criteria for judging test performance which are reflective of the indigenous assessment practices of these domain experts in the pertinent occupational or workplace setting. While recent LSP research has focused on the development of assessment criteria that closely reflect the values that underpin domain experts’ assessment of performance in the target domain, little attention has been paid to a later stage in the testing process – setting standards on a test – where domain experts tend also to be involved in determining how much proficiency is enough for the relevant occupational or workplace domain. All of the above issues are captured in the OET, considered to be an exemplar of an LSP test, but also reflecting many of the complexities and uncertainties surrounding LSP research and practice including the scant attention to standard setting, mentioned above. The next section turns to a discussion of standard setting in language assessment.

2.5 Standard setting

This section firstly presents background on standard setting. Then, a categorization of a range of standard-setting methods is provided, specific criteria for selecting a
method are stated, and an assessment of a number of potential methods for use in the present study is offered. Following that, several standard setting reports and studies (particularly language testing and health profession studies) are mentioned, including recent standard-setting studies on the OET. Possible concerns in standard-setting research are acknowledged such as variability, method effects and domain expert involvement. Finally, the value of qualitative methods in standard-setting research as an instrument for investigating the validity of participants’ judgements is emphasized.

### 2.5.1 Background

Standard setting concerns the use of socially mediated methods to connect test performances to more generalizable explanations of those performances (Cizek, 2001). These explanations are about what a score means in the context of the real-world performance the test is designed to predict – for example, the explanation might be to ascertain the effect that achieving a particular score on a test indicates readiness to operate in a particular profession or preparedness for training – and by implication the failure to achieve that threshold score is an indication of insufficiency in the ability or skill set under assessment. Furthermore, standard setting is a practice that attempts to follow a defined procedure using qualified participants (e.g., domain experts with knowledge of the context in which test takers are likely to perform) to give interpretive significance to those performances while trying to diminish arbitrariness (Kenyon & Römhild, 2014).

The setting of performance standards is a potentially challenging matter confronting policy makers and language test developers. Cizek (2012) emphasises the importance of standard setting in test development as “it is difficult to imagine a higher-profile, more consequential, and more contested element than the setting of performance standards” (p. ix). Bachman and Purpura (2008) note that language assessments, and particularly high-stakes ones, can be classed as ‘gatekeepers’ or ‘door-openers’ in the sense that attaining or failing to achieve a passing standard will impact the life chances of test candidates either by closing off opportunities or creating avenues for
progression. Kane, Crooks and Cohen (1999a) observe that on tests such as licensing and certification, the setting of passing standards can have two sides: the needs of the public and the needs of test candidates. On one side, health and safety may be affected if test candidates’ proficiency and skills are not at a high enough standard. This is especially critical in professions such as aviation or medicine because deciding if a candidate is equipped to perform their profession or not may be crucial and sometimes even a life-threatening question. On the other side, professionals should be allowed to work in the field for which they have the necessary experience and qualifications. Capable test candidates may be rejected, or their numbers limited if the standard is set too high. Standard setters therefore need to ensure that decisions reached on the basis of test scores are valid and founded on mutual awareness of how much knowledge, skills or language proficiency (as the case may be) is satisfactory to perform in a certain domain. The central concern is validity (i.e. the defensibility of the inferences generated from test scores because individual arbitrators can differ in the basis for their judgements). Thus, to diminish arbitrariness in decision making, it is critical to adopt a principled procedure for reaching well-founded consensus on appropriate levels of performance.

Procedural consistency is a key concern which is often stressed in the standard-setting literature and it can be further defined as the “proper following of a prescribed, rational system of rules or procedures resulting in the assignment of a number to differentiate between two or more conceivable states or degrees of performance” (Cizek, 1993, p. 100). Tannenbaum and Cho (2014) state that standard setting “refers to a variety of systematic, judgement-based processes that identify a minimum test score that separates one level of performance (e.g., proficiency, understanding, or expertise) from another” (p. 235). These processes importantly utilize experienced and knowledgeable panellists to arrive at independent judgements and through consensus-building procedures, record discrepancies and indicate the degree to which agreement was achieved (Kenyon & Römhild, 2014). Judges are asked to decide where to ‘draw the line’ which is independent of the regular test marking process (Kane, 1994). Setting a passing standard or ‘cut score’ between one performance level and another is one of the final yet fundamentally significant steps in the procedure of
guaranteeing that defensible evaluations are made from test scores. A cut score is a place on a test’s score range that is used to establish if a certain assessment score is adequate or not for some purpose (Zieky et al., 2008). Statistical calculations are often used to determine what the cut scores should be for judging who passes and fails a test or who is awarded a particular grade. Cizek and Bunch (2007) characterize standard setting as a systematic procedure where a concept (i.e. the performance standard) is interpreted as an operational position on a score scale (i.e. the cut score). In other words, a cut score is the ‘operational version’ of a required performance level and the ‘performance standard’ is the conceptualization of the required performance level (Papageorgiou, 2010). For instance, on the OET Writing sub-test, there is a cut score indicating the operational position on the score scale for each of five grade levels (A, B, C, D or E) with the B grade performance standard defined as “high level of performance i.e. able to use English with fluency and accuracy adequate for professional needs” (OET, 2017).

Zieky et al. (2008) stress, however, that a ‘perfect’ cut score is an illusion and not the intended aim of standard-setting study. A cut score is always a value judgement with the intention being to ensure that it is an informed and defensible assessment elicited from people whose opinion counts in the context of concern. Cut scores are determined by the people who set them and a number of differing variables (e.g., the beliefs and experience of the participants, the specific method used to set the cut score and the manner in which a chosen method is applied by a facilitator). Any alteration in these aspects during the process may result in another cut score outcome. Indeed, as Hambleton and Pitoniak (2006) point out, there is no clear agreement on the preferred option of standard-setting methods from the many alternatives available. They refer to this uncertain situation as the ‘Achilles heel’ of educational testing. Likewise, Kaftandjieva (2004) states that the ‘one best’ method for standard setting is yet to exist, but the grounds for the chosen procedure need to be justified. Some of the best-documented methods related to constructed performances (of the kind that are assessed on the OET Writing sub-test) are discussed in the following sections.
2.5.2 Categorization of standard-setting methods

There has been considerable research on standard setting undertaken in the fields of psychometrics and educational measurement (see e.g., Angoff, 1971; Cizek, 2001, 2012; Cizek & Bunch, 2007; Ebel, 1972; Jaeger, 1989; Thurstone, 1927; Zieky et al., 2008). In her exhaustive review, Kaftandjieva (2004) notes that as many as 50 standard-setting methodologies have been documented and a range of adaptations are present for a large number of them. These methods are often further classified and Jaeger’s (1989) categorization system of separating methods into ‘test-centred’ and ‘examinee-centred’ types is one that is frequently used. Test-centred methods focus on the judgements made about test tasks or items. In these methods, panellists evaluate individual test items or questions to single out the performance limit between ‘minimally competent’ and ‘incompetent’ of an ‘envisaged’ or ‘imagined’ test taker (Cizek & Bunch, 2007). Cizek (1996) notes that for item-based methods, the ‘Angoff’ and variations such as the ‘modified’ and ‘extended Angoff’ are the most extensively used and researched procedures. Zieky et al. (2008) state that in these types of test item-centred methods, panellists’ “judgements of test questions are based on the idea that an appropriate cut score would be the test score expected of a typical borderline test taker, one whose knowledge and skills are just barely good enough to be included in a proficiency level” (p. 61). In test-centred methods, the standard-setting judgement is focused on test tasks/items and a theoretical or supposed candidate’s ability to answer those tasks or items correctly (or not).

In examinee-centred methods, holistic judgements are made about ‘real’ examinees and/or their products of assessment (Cizek & Bunch, 2007). The focal point is on ‘actual’ candidates and/or their performances so the judgement is more direct and less hypothetical. Haertel and Loriè (2000) further distinguish an area within the examinee-centred category to centre on examinee performance only called ‘performance-centred.’ This is because some examinee-centred methods such as the ‘Borderline’ and ‘Contrasting Groups’ initially focus on test takers themselves rather than their performances. Kaftandjieva (2004) notes in addition to Jaeger’s (1989) two category system of ‘test-centred’ and ‘examinee-centred’ types, another third overall
group named ‘other methods’ is sometimes used. This includes a narrow range of methods that focus not on test items or test taker performances, but on test candidates’ score distributions or score profiles. Cut scores can then be made on judgements of these score distributions or profiles and through the use of empirical data.

Kaftandjieva (2004) observes that due to the constraints of Jaeger’s binary scheme, Hambleton, Jaeger, Plake and Mills (2000) and Reckase (2000) offer their own classification schemes that consider more fine-grained characteristics of the standard setting process. Hambleton et al.’s (2000) categorization system centres on: the stimulus for panellists’ decisions (i.e. test takers, tasks, work results, rated performances); the decision task given to the panel; the judgemental procedure; the arrangement and number of panel members; the justification of the subsequent standards; and the type of test task. Furthermore, Reckase (2000) proposes three categorizations: the size or complexity of the judgement task; the amount and type of supporting information and feedback provided to judges; and the complexity of the method applied for cut score establishment. While Jager’s (1989) two-category system focussing on the judgement remains the most commonly employed one, it leaves a lot of considerations unanswered. It seems clear that additional aspects of the standard-setting process need to be considered when choosing a method. The following section highlights some key considerations when selecting a method for a standard-setting study.

2.5.3 Selecting a method

Kaftandjieva (2004) notes that of the many standard-setting methods that have been documented, each has its individual advantages and disadvantages. When deciding on a method, several authors recommend a variety of selection criteria, but the central ones include: the suitability of the method for the specific context; the viability of the method’s procedure in the present situation; number of items; sample size; stakes (high or low); perceptions and/or evidence about the validity of the chosen method; and available resources in terms of time, staff, funding, equipment, degree of
expertise, and software available (Cizek, 1996; Kaftandjieva, 2004; Reckase, 2000 Zieky et al., 2008).

Cizek (1996) states that the integrity of cut scores generated from a standard-setting study would be strengthened if there were sufficient preceding verification of the quality of the method (i.e. defensible cut scores require a solid and well-researched method). Moreover, Norcini and Shea (1997) add that “standard-setting methods that require effort are likely to be viewed as more credible than those that do not” (p. 44). Kaftandjieva (2004), however, cautions that “the intent is to demonstrate due diligence, not endurance” (p. 44) pointing out real-world practical restrictions are of significant consequence and cannot be ignored in choosing a method. It nevertheless remains important that “if a less widespread standard-setting method is preferred, then a detailed methodological description of the method should be provided together with sound and compelling arguments for its development and implementation” (Kaftandjieva, 2004, p. 28).

Berk (1986) suggests two overarching themes for assessing a standard-setting method to be used in a study: technical adequacy and practicability. A technically adequate method generates suitable classification data (cut scores), is sensitive to instruction or training, is sensitive to examinee performance, ascertains the true standard (i.e. allows for measurement error), is statistically rigorous and generates decision validity evidence (i.e. offers approximation of decision consistency). A practical method is one that is reliable and also easy to implement, compute and translate to the general public. More recently, Hambleton and Pitoniak (2006) encompassed Berk’s (1986) two general themes and proposed four key factors that need to be considered when selecting a suitable standard-setting method. Firstly, the test’s item arrangement needs to be taken into account (i.e. whether the test comprises multiple-choice items only) (Angoff, 1971; Nedelsky, 1954), performance assessments (Cizek & Bunch, 2007) and mixed item assessments or score profiles (Hambleton et al., 2000; Kingston & Tiemann, 2012). Secondly, when conducting a standard-setting study, time and resources may be restricted, and therefore it is necessary to align the time required to apply a method. Thirdly, as noted by Kaftandjieva (2004), because trialling a new
method or modification can be time-consuming and expensive a previously used procedure should be borne in mind. Lastly, the validity of a method needs to be explored (Bahry, Hambleton, Gotzmann, De Champlain, & Roy, 2012).

It should be noted here that the OET comprises a mix of tasks/question items depending on the skill area being assessed. However, the OET Writing sub-test, which is the focus of the current research, is a single item performance test for which only some standard-setting methods may be suitable (see below). Furthermore, the relevant medical domain experts to be involved in this study’s standard-setting procedure are often employed in high pressure and time-poor environments and may have a limited schedule or commitment available to participate in a lengthy and elaborate standard-setting study. The next section presents some of the common standard-setting methods used in assessing constructed performances such as writing responses and note some of their advantages and drawbacks in relation to being used as a basis for setting cut scores on the OET Writing sub-test.

2.5.4 Evaluation of different methods

In language testing contexts, there are several methods for setting cut scores that rely on judgements of the work or performances that test takers produce. For these methods, standard setters need to be able to assign each test candidate’s skills and knowledge to a performance level to determine final cut scores (Zieky et al., 2008). Many of these standard-setting methods’ main approach is to appraise a range of language performance samples and usually have a panel of judges give a single (holistic) decision about each sample. Judges make holistic judgements of the performances and consider whether a candidate is in a ‘borderline’ or ‘just qualified’ range such as ‘pass’ or ‘fail’ or ‘proficient’ or ‘basic’ and these areas are used to set cut scores (Tannenbaum & Katz, 2013). Some of these methods also attempt to link a performance test score to an account of language skills or an operational description that expresses a scale of language skill levels or competencies, often in the form of ‘can do’ statements: the foremost of these in recent times has been the Common European Framework of Reference (CEFR) (Council of Europe, 2011; Tannenbaum &
Cho, 2014). Through this process, meaning can be given to scores “by referencing them to . . . performance levels, benchmark performance levels, or achievement levels” (Kane, 2012b, p. 8).

A variety of standard-setting methods are described in the literature for setting standards on performance-based tasks such as writing (see e.g., Cizek, 2001, 2012; Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006; Kaftandjieva, 2004; Zieky et al., 2008). As noted, a range of examinee-centred methods require standard-setting panellists to make overall judgements about assessment performances (Cizek & Bunch, 2007) and require participants to consider if a test candidate is in a ‘borderline’ or ‘just qualified’ range to set cut scores (Tannenbaum & Katz, 2013). A benefit of these approaches is that it is arguably easier for assessors to make a determination about an actual test-taker (Hambleton & Pitoniak, 2006; Kingston et al., 2001) than to estimate the likely performance of a hypothetical group of test candidates on particular test items such as with test-centred methods. Cizek (1996) argues that the former type of judgement process is more ‘natural’ or less contrived than making judgements about item content. These examinee-centred and/or performance-centred methods may also be more suited to constructed-response items such as the OET Writing sub-test task. They may also be more manageable for domain experts who might have even more difficulty than language experts in making judgements about particular test items as required in test-centred approaches. However, the consequences of involving domain experts who are not necessarily trained in the area of language proficiency (the area they are being asked to evaluate and differentiate from clinical competence/knowledge) may be an issue for any of the standard-setting methods being considered (see Pill & McNamara, 2016). The capacity of domain experts to make such evaluations in a defensible manner is explored in the current study.

The following sections (2.5.4.1 to 2.5.4.7) examine several standard-setting methods for possible use in this study with constructed test candidate responses such as the OET Writing sub-test task. The first set of methods are examinee-centred and are further divided as follows: 1) a focus on examinees themselves – the ‘Contrasting
Groups’ and ‘Borderline Group’ methods; 2) a focus on examinees’ constructed responses – the ‘Body of Work’ method, the ‘Analytic or Analytical Judgement’ method, the ‘Performance Profile’ method, and the ‘Examinee Paper Selection’ method. The second set of methods are test-centred, but they have also been previously used in studies with examinees’ constructed responses: the ‘Angoff’ method and the ‘Bookmark’ method.

2.5.4.1 The Contrasting Groups and Borderline Group methods

Two similar approaches that have often been employed for standard-setting practices with constructed performances are the ‘Contrasting Groups’ (Berk, 1976) and ‘Borderline Group’ (Livingston & Zieky, 1982) methods that centre on the appraisal of real test takers known to standard-setting panellists (Kenyon & Römhild, 2014). The Contrasting Groups method separates test candidates at each score level into two contrasting sets on the basis of judgements of test takers’ knowledge and skills: a group that is qualified to belong in a performance level and one that is not. Test takers could for example be divided into groups that are ‘Proficient’ and ‘Basic’ respectively (Livingston & Zieky, 1982). Zieky et al. (2008) state that “one reasonable choice for a cut score would be the score at which 50% of the essays are ‘Proficient’ as that would represent the borderline of the Proficient performance level” (p. 78). Similar procedures could be applied for other cut score boundaries (i.e. between ‘Proficient’ or ‘Advanced’). Benefits of this method are that it is relatively easy to explain and large sample sizes can be judged by the same group of trained participants (Zieky et al., 2008).

The Borderline Group is a similar method and is founded on the concept that when a test candidate’s performance is on the borderline of a performance, the cut score should be allocated from around this candidate’s actual test score (Schoon & Smith, 1996; Searle, 2000). According to Zieky et al. (2008), assessors do not have to evaluate all the test candidates or a representative sample, but only the ones considered as borderline. A cut score would be set at the median score (the 50th percentile) of the
borderline group. They argue that for the simplicity of its use it is one of this method’s main advantages, however it may be difficult to identify truly borderline candidates.

An issue with both the Contrasting Groups and Borderline Group methods is that they were developed in the “era of awakening (1960-1980)” and their usual implementation is to ask panellists to make judgements for the entire instructional timeframe about ‘known’ test takers just as teachers would do with their own class of students (Kaftandjieva, 2004, p. 15). Also, judgements have been conventionally made on information outside of the assessment period for which performance standards are to be established and consequently potential bias could affect standard setters’ judgements such as student likability and effort (Cohen, Kane, & Crooks, 1999). This compares unfavourably to more recently developed methods where “the judgements about each examinee are based only on his/her overall performance on the test under consideration” (Kaftandjieva, 2004, p. 15). Accordingly, as the authors (i.e. test takers) of the OET writing samples are unknown to the standard-setting participants in this study, other examinee performance-centred methods reviewed below may be more suitable.

2.5.4.2 The Body of Work method

The ‘Body of Work’ (BOW) method was developed to evaluate whole response collections from test takers (Cizek & Bunch, 2007; Kingston, Kahl, Sweeney, & Bay, 2001; Kingston & Tiemann, 2012) and centres on classifying student work instead of the students themselves. It is often known as the ‘holistic’ method (Hambleton et al., 2000), a term which also somewhat confusingly applies to the umbrella family of methods that the BOW is a part of (Hambleton & Pitoniak, 2006). The main feature of the BOW is the appraisal of a complete set or sub-sets of test taker performances (although the method could also potentially be applied to a single constructed response such as OET writing samples as panellists still give an overall, holistic decision). The judgement task in the BOW is to ascertain and match the demonstrated knowledge and skills shown in an examinee response to a corresponding performance level (Cizek & Bunch, 2007). The overall approach of the BOW is to evaluate student
work samples in multiple rounds where panellists make progressively more specific judgements and classifications. Work samples are arranged in folders by total score from high to low or low to high, but panellists do not know the exact score for any sample. Kenyon and Römhild (2014) state that the BOW is a process that can be applied repeatedly starting with a ‘range-finding’ step carried out by panellists where work samples are looked at roughly to decide where an approximation of cut scores might be placed. This is achieved by evaluating samples at quite broadly spaced score positions and deciding the general performance level classification within which each sample could be categorized. Then, panellists conduct a more detailed process known as ‘pinpointing’ that reviews additional examinee responses within narrower ranges of the initial estimated cut score areas. More scripts are provided exemplifying the different score points within these approximated score ranges in order to establish a more exact location for the cut score/s (Kenyon & Römhild, 2014).

Hambleton and Pitoniak (2006) argue that this two-step process makes examinee response review more manageable because response sets from non-critical score areas (i.e. those that are far away from the relevant cut score) can be eliminated to allow extra samples from critical score areas to be included and the procedure of attaining more judgements closer to the cut score regions yields increased reliability of the ultimate cut scores. Final cut scores are usually established through analytic practices such as logistic regression that “models the relationship between a continuous variable, such as a test score, and the probability of being in a binary category, such as being judged as being proficient or above” (Kingston et al., 2001, p. 230).

A benefit of the approach’s judgement task is that it is easier for examiners who are familiar with the task of assessing constructed performances as it is arguably more intuitive than a test-centered method that asks panellists to imagine the probable performance of a hypothetical test taker (Hambleton & Pitoniak, 2006; Kingston et al., 2001). However, for the health-related domain experts involved in setting standards on the OET, evaluating and assessing language test samples produced by second language learners may not be easy, given their lack of experience in this area. One
concern with the BOW is that in relation to other standard-setting methods the cut scores derived from this approach are sometimes higher or more stringent than those derived from other methods (Hambleton & Pitoniak, 2006). Another issue is the time involved in judging performances and a resultant fatigue effect which may limit the amount of examinee responses that panellists can validly and consistently assess in a session (Zieky et al., 2008). In addition, there may be questions to consider about which examinee samples are selected and the order of their arrangement in folders.

2.5.4.3 The Analytic or Analytical Judgement method

Another approach from the holistic group of methods is the ‘Analytic or Analytical Judgement’ method (AJM) (Plake & Hambleton, 2001). This methodology is similar to the Contrasting Groups and Borderline Group procedures in that it examines borderline performances, however it can be used with ‘unknown’ test candidates. The method also has the advantage of being a multiple borderline grouping method (Cizek & Bunch, 2007), meaning that multiple cut scores can be set in the same round of judgement such that candidates are assigned to different performance levels. The AJM uses, for example, four performance levels such as ‘Advanced’, ‘Proficient’, ‘Apprentice’ and ‘Novice’ (similar to the A, B, C and D grades on the OET) each of which can be further sub-divided into ‘High’, ‘Medium’ and ‘Low’ to make 12 groupings overall (Plake & Hambleton, 2001). Participants are initially invited to examine and discuss each of the main performance categories for which judgements are decided (e.g. panellists are directed to consider what features would make a test candidate’s writing response ‘Proficient’ or ‘Apprentice’ level). To arrive at cut scores, assessors individually review assessment samples and categorise them according to each of the 12 sub-divided performance levels. Standard-setting participants’ judgements are then tallied and combined into further new groupings around the boundary categories (i.e. those that are just either side of four performance levels such as ‘High Basic’ and ‘Low Proficient’). The average of the total scores for each new boundary category is calculated to give a final cut score determination. From Plake and Hambleton’s (2001) own three field tests, they concluded that for ease of use reasons the method is very beneficial. However, Cizek and Bunch (2007) and Plake and Hambleton (2001)
themselves suggest that although the 12-category system works well it could be simplified into a 7-category method by eliminating the ‘medium’ category in each performance level. Nevertheless, it could be argued that the original 12-point classification allows greater accuracy and narrowing of panellists’ focus in achieving performance level judgements.

In its intended implementation, the AJM also utilises external experts to rate the assessment samples and according to Cizek and Bunch (2007), this distinguishes it from other holistic procedures such as Contrasting Groups and Borderline Group methods as these merely rely on their own teacher’s evaluation of their students’ performance and not on outside expert professional judgement. They argue that this increased professional precision from subject-area experts gives greater credibility to this standard-setting methodology. However, Zieky et al. (2008) note that the method can also be time-consuming for panellists due to the number of samples that need to be assessed and the requirement to classify the samples within a performance level. This could be alleviated though by giving participants samples to classify after the initial standard-setting workshop to assess in their own time or by employing a crossed design in which different participants assess different samples.

2.5.4.4 The Performance Profile method

An additional examinee-centred method that can be used for constructed-response tests is the ‘Performance Profile’ method. The ‘profile’ is the test candidate’s pattern of scores on individual parts of the test and the method is useful for assessments that consist of multiple writing task responses (Zieky et al., 2008). In a training session, participants are asked to discuss what aspects would make a test candidate’s writing response ‘minimally competent’ or ‘borderline’. The writing samples are then presented in a binder or ‘Ordered Profile Booklet’ (OPB) for the panellists in increasing total-score order. Participants consider the ordered profiles and choose one that is suggestive of borderline performance at a stated performance level (e.g., between pass/fail or B/C on the OET). Panellists then look at all the profiles with the same total score as the first chosen borderline profile (i.e. the first borderline profile may have
had scores of 4, 6, 3, 5 with a total of 18 and the next profile at 9, 2, 1, 6). If all the profiles at the chosen total score are assessed to represent ‘Proficient’ performance, the decided total score is the panellist’s cut score for that particular performance level. The mean, median or trimmed mean can be used to obtain the groups’ overall cut score for each performance level.

One of the advantages of the method is that, like the Analytic Judgement method, it can be used to set multiple performance levels in the same round. On the other hand, significant time is needed to prepare profiles at each score level to create the OPBs, making the method potentially unwieldy if there are more than seven scored questions. The method is also quite new and not as well researched as others (Zieky et al., 2008). In addition, importantly for the current study, the method is designed, like the Bookmark method described below, to be used on tests with multiple writing questions/tasks which are then ordered in level of difficulty. The OET Writing assessment is a single question task and therefore other methods may be more appropriate. However, as the overall judgement process and setting of cut scores is similar when assessing a single sample or a group of samples (profile) (as also seen in the BOW method), a modified version of the Performance Profile method as used by Tannenbaum and Wylie (2013) could also be applied to the OET Writing sub-test task. Their variation of a Performance Profile approach for the Test of English for International Communication (TOEIC) Writing test used eight different task scores. The responses were presented in a folder for the participants in increasing total score order and participants evaluated the performance profiles and judged these at CEFR approximated levels. This method could be adapted for use in this study’s standard-setting procedure.

2.5.4.5 The Examinee Paper Selection method

Another performance sample ordering method is the ‘Examinee Paper Selection’ method (Hambleton, et al., 2000) that orders test takers from worst to best or best to worst. It is similar to the Borderline Group method; however, participants evaluate actual examinee test samples instead of using their independent knowledge and
rankings of the test candidates. Participants choose from already graded test writing samples to consider those that are representative of a borderline candidate (e.g., between pass/fail or OET band B/C). Panellists are told to choose one or more papers to demonstrate performance at the borderline of each achievement level. The score allocated to the sample is not shown to the participants, though “the score may be ascertained through logical deductions from the feedback information provided between three rating rounds” (Loomis & Bourque, 2001, p. 183). “The average of the scores on the selected papers for each performance standard is used as the ‘minimum passing value’ for each question” and “these minimum values are summed to determine each performance standard” (Hambleton, et al., 2000, p. 359).

This type of constructed response method in its traditional implementation would have to be modified for the planned OET study as it only sets a ‘pass’ or ‘fail’ cut score and not multiple band levels as on the OET. For example, participants could decide on the sample/s that are consistent with their idea of a ‘minimally competent candidate’ (MCC) (i.e. grade B on the OET), and then make further decisions about the grade A (highly competent), as well as the less than competent C, D and E levels. Hambleton et al. (2000) assert that participants (especially teachers) value the method because it bases their decisions on actual test candidate samples and ranking papers is a common task, but the method is not easy to realize with small test candidate numbers and it may be difficult to obtain representative samples of each performance level from the distribution of test candidate work. Hence, other methods may be more suited to this study’s standard-setting procedure.

2.5.4.6 The Angoff method

One test-centred approach that is sometimes used with constructed responses is the ‘Angoff’ method (Angoff, 1971) or ‘modified/extended Angoff’ method. A traditionally implemented Angoff method asks assessors to evaluate individual test items as likely to be scored either ‘correct’ or ‘incorrect’ by a ‘minimally competent candidate’ (MCC). Plake and Cizek (2012) assert that the method could also be applied to contexts where more than a dichotomous pass/fail score is required with multiple cut score
regions (such as on the OET). The method was originally designed for use with multiple choice question (MCQ) type tests, but it has been used with constructed response items (see Qian, Woo, & Banerjee, 2014), and “the procedure would again be to determine how minimally competent, or borderline candidates will likely perform on the task” and “cut scores for the overall test are obtained by aggregating the expected performances across all the tasks that comprise the test” (Plake & Cizek, 2012, p. 192).

In a variation, the ‘Angoff Mean Estimation’ method, Zieky et al. (2008) note that “instead of asking participants to state the probability that borderline test takers would get the question right, simply ask the participants to estimate the average score that a large group of borderline test takers would obtain of the question” (p. 63). Plake and Cizek (2012) state “panellists would estimate the mean score that a MCC would earn on the constructed response tasks. These mean item estimations can then be averaged across panellists to get the overall estimate of the mean performance of MCCs at a particular cut score for the task” (p. 192). A modified version of the method has been used (see Qian et al., 2014 in the next section) to compare panellist’s borderline judgements of real test takers to sample candidate papers on a writing assessment comprising multiple constructed response tasks/questions. A main criticism of the method is the difficulty of the participants’ task of formulating item performance estimates and of making “consistent and accurate probability estimations” (Plake & Cizek, 2012, p. 192), therefore other methods may be more appropriate.

### 2.5.4.7 The Bookmark method

The ‘Bookmark’ method (Lewis, Mitzel, & Green, 1996; Lewis, Green, Mitzel, Baum, & Patz, 1998; Lewis, Mitzel, Mercado, & Schultz, 2012; Mitzel, Lewis, Patz, & Green, 2001) is a test-centred standard-setting procedure that is often used with individual test items in Reading and Listening assessments although it can be used with constructed-response items. In part, the method was developed due to criticism of the Angoff method as panellists found the task of approximating item difficulty
precisely for borderline examinees to be challenging (Goodwin, 1999; Impara & Plake, 1998). The bookmark method is meant to offset this demand by requiring panellists to study a booklet of test items organized in order of difficulty and make a decision about where in this ordering the proposed standard rests (Skaggs, Hein, & Awuor, 2007). Kenyon and Römhild (2014) note the method in its traditional utilization employs item-mapping procedures “to convey through spatial representations the relationship between item content, item difficulty, and the measurement scale on which cut scores are set” (p. 7). Test scores are regularly standardized with ‘Item Response Theory’ (IRT) prior to the standard-setting session. Test items or questions are shown in degree of difficulty from easy to hard in a special test booklet called an ‘Ordered Item Booklet’ (OIB). Panellists are required to put a ‘bookmark’ at the position that divides the items into groupings (e.g., Proficient/Basic). The implied cut score for a borderline test candidate can be ascertained when a panellist has completed their bookmark placement. The overall cut score is established as the median bookmark placement of panellists (Lewis et al., 2012; Zieky et al., 2008).

Zieky et al., (2008) note that the efficiency of the method is a particular advantage as multiple cut scores can be set during a standard-setting session on the same test and the method is suitable for constructed-response items. For the OET Writing sub-tests, instead of test question items ordered in level of difficulty, samples could be ordered from low to high or high to low, without informing candidates of the precise score for any work sample. This ordering of samples is similar to the arrangement of folders in the Body of Work (BOW) method, however in this method’s operation there would be fewer rounds of ‘range finding’ and ‘pinpointing’. Zieky et al. (2008) observe that time-consuming procedural issues may be hindrances with the method such as producing the OIBs can be protracted and IRT-calibration of all the test items needs to be completed prior to the cut score session. This method is traditionally a question-judgement oriented procedure to be used on tests with multiple writing questions that orders them in level of difficulty. In regard to the OET, the Writing assessment is a single question task, but it could be adapted to fit this method in a ‘Simplified Bookmark Method’ (Karantonis & Sireci, 2006).
To sum up, the standard-setting methods that have been evaluated in the preceding section have their own positive and negative points in relation to a constructed performance test such as the OET Writing sub-test. As noted previously, when selecting a suitable method for a standard-setting study a number of selection criteria need to be taken into account such as suitability for the specific context, viability, sample size, stakes (high or low), perceptions and/or evidence of validity and available resources (Cizek, 1996; Kaftandjieva, 2004; Reckase, 2000; Zieky et al., 2008). From the prior evaluation of several methods, two of the most suitable for this study were chosen to be trialled in order for one final method to be selected: the Analytic or Analytical Judgement method and the Performance Profile method. The reasons for the choice of these two trial methods and the results of two pilot studies are outlined and discussed in the method chapter.

2.5.5 Studies and reports on standard setting

The following section firstly notes studies related to language testing and standard setting that are linked to language proficiency frameworks. Then, more specific standard-setting research related to the health professions are highlighted. Lastly, standard-setting research on the OET is described.

2.5.5.1 Language testing and standard setting

In language testing contexts, issues associated with standard setting are attracting increasing interest among researchers, although studies documenting standard-setting practices are still relatively uncommon. Recent standard-setting studies with a language testing focus have, in the main, been conducted to offer greater understanding of test scores by linking them to explanations or operational definitions of language skills (Tannenbaum & Wylie, 2013). Some studies have sought to link specific assessments to the Common European Framework of Reference (CEFR) (see e.g., Dunlea & Matsudaira, 2009; O’Sullivan, 2012; Papageorgiou, 2007; Papageorgiou & Tannenbaum, 2016; Tannenbaum & Wylie, 2004, 2005, 2008, 2013; Taylor, 2007). In this situation, the primary aim of standard setting is to “identify minimally
acceptable scores (cut scores) needed to enter the different CEFR levels of interest” (Tannenbaum & Wylie, 2013, p. 1). These procedures are not substantially dissimilar from other standard-setting approaches, but the main distinction is that the cut score level being set on a particular test is linked to a level on another scale (the CEFR), rather than being tied to more absolute notions such as ‘minimally competent’ or ‘proficient’. As the focus of this thesis is on health communication and the OET, the following section highlights several key standard setting studies about language proficiency conducted with the health professions.

2.5.5.2 Health professions and standard setting

In the medical education literature, standard setting has received much attention (see e.g., Bahry et al., 2012; Cusimano, 1996; Norcini, 2003). The initial form of many of the standard-setting methods was in written, multiple-choice assessment designs (Boursicot, 2006). Some of these standard-setting procedures have now been employed with complex performance-based assessments such as the Objective Structured Clinical Examination (OSCE) (in a number of jurisdictions in the United States) that tests professional knowledge and performance skills competence as well as professional communication in a clinical examination. Many of the OSCE tests use test-centred standard-setting methods such as the Angoff method or examinee-centred methods such as the Borderline Group method depending on the skill being assessed. However, relatively little standard-setting research has paid exclusive attention to written communicative proficiency and its function in the management of care of patients (Boulet et al., 2004).

In health professional contexts that focus on language, several standard-setting studies are evident, although they are limited in number. Notable examples include studies employed by the United States National Council of State Boards of Nursing (NCSBN) in establishing the passing standards of foreign-trained nurses who seek to gain licensing and certification in the United States. These studies used a variety of national and international English language proficiency tests (e.g., IELTS and TOEFL) and standard-setting methods (see below) to indicate the required entry level for
graduate nurses to perform safely and effectively. In many of these studies a diverse range of domain experts were employed as panellists such as registered nurses, nurse educators, nurse regulators, nurses who regularly engage with English as a Second Language (ESL) clients, nurses with ESL and clinical nurse supervisors with ESL.

O’Neill (2004), O’Neill, Marks and Wendt (2005) asked subject specialists in the domain of nursing to make judgements and recommend a passing cut score on the paper-based Test of English as a Foreign Language (TOEFL). These studies used two standard-setting methods in establishing a passing level for a ‘minimally competent examinee’: The ‘Simulated Minimally Competent Examinee’ (SMCC) method (a modified Angoff method) and the Examinee Paper Selection method (described above). Also, Wendt and Woo (2009), using this same method and a panel of nursing experts, compared the TOEFL internet-based test (iBT), with prior nursing passing standards recommended for TOEFL and IELTS. They found that the iBT version was consistent with previous standards for nurses and the required level of English proficiency.

Woo, Dickison and de Jong (2010) set passing standards with a panel of subject-matter experts in the field of nursing on the Pearson Test of English Academic, an internet-based test that is aligned with the Common European Framework of Reference (CEFR). For the Writing sub-test, using the ‘Performance Profile’ method (described above), panellists reviewed response profiles individually from actual test candidates at a range of score levels (i.e. ‘Basic’ to ‘Proficient’ English users). In two rounds of ratings, panellists initially estimated the ‘absolute ability’ of test takers in relation to the descriptors in the CEFR. Secondly, participants approximated the ability of entry-level nurses for a minimum level of English language proficiency. Overall, a single cut score of minimum English proficiency for the Writing sub-test was recommended through group discussion and an average of panellists’ cut scores.

In another study by Qian et al. (2014), domain expert panellists in nursing used an extended Angoff method to independently review test-takers’ performances for the Writing sub-test at numerous score levels (from Basic to Proficient English users) on
the Michigan English Language Assessment Battery. In this version of an extended Angoff method, instead of estimating the percentage of minimally competent test takers who will accurately answer a question, participants estimated the amount of points minimally competent test takers would need to attain on each question using a set of predetermined criteria. The standard-setting assessors for the Writing sub-test were given a collection of test candidate responses at a range of score profiles on the rating rubric (from Basic to Proficient English user). Panellists evaluated the responses individually and conferred in a panel discussion on the question of whether an entry-level nurse with ‘minimally acceptable’ English proficiency would be able to answer questions similar to those of the sample candidates. This process, involving two rounds of rating and discussion, yielded consensus on the cut score that an entry-level nurse would need to achieve to perform effectively and safely in the workplace.

Standard-setting studies have also focused on the use of IELTS for a range of health professions in different jurisdictions. O’Neill, Buckendahl, Plake and Taylor (2007) used a modified Analytic Judgement method (AJM) (described above) with the writing performances to set a passing standard for nurses in the United States with a panel of subject-matter experts. Firstly, judges singled out three writing samples that were the worst of the ‘Competent’ samples and the three best of the ‘Incompetent’ samples which were averaged to produce a round one cut score. Judges were provided with feedback data on the mean of the panellists’ cut scores allowing them to review and change their original individual scores. This resulted in the round two cut score. Because the IELTS writing test has two tasks, the “final band score was calculated using the panellists’ cut score for each task and multiplying it by the weighting for each task and summing those scores” (O’Neill et al., 2007, p. 309).

In another similar study, but in a different jurisdiction, Berry, O’Sullivan and Rugea (2013), in a report to the General Medical Council (GMC) in the UK, recommended new standards of IELTS scores for IMG doctors. This standard-setting panel involved distinct groups of doctors, nurses, allied health professionals, medical directors and patients/members of the public. Writing samples were evaluated using the Examinee Paper Selection method (Hambleton, et al., 2000) (described above) that ranks test
takers from worst to best. After a number of rounds, panellists chose the sample that displayed the minimum satisfactory level of writing for a ‘minimally competent candidate’ (MCC). The final recommendation suggested a higher passing standard than that already in place.

The majority of the tests that were the focus of the above-mentioned standard-setting procedures were general proficiency tests, rather than profession-specific language (LSP) tests like the OET that draw on professionally relevant materials and use tasks designed to simulate the language demands of that profession. However, domain experts were involved in all of the previously reviewed health-related standard-setting exercises and subject-matter specialist inclusion was considered vital (as in most standard-setting procedures). Studies of standard-setting efforts in relation to profession-specific LSP tests are extremely rare. An exception to the lack of an LSP test in the health professions (apart from the OET) is the Canadian English Language Benchmark Assessment for Nurses (CELBAN) (2015), purposely designed for foreign-trained nurses. This test uses the Canadian Language Benchmarks (CLB) as the standard to determine the language proficiency demands of nurses and employs a 12-point “descriptive scale of communicative proficiency in ESL, expressed as benchmarks or reference points” (CELBAN, 2015, p. iii). These are expressed in three stages: “Stage I, Basic Proficiency (Levels 1 to 4); Stage II, Intermediate Proficiency (Levels 5 to 8); and Stage III, Advanced Proficiency (Levels 9 to 12)” (CELBAN, 2015, p. ii). The descriptors are similar to those found in the Common European Framework of Reference for Languages (CEFR) and in the American Council on the Teaching of Foreign Languages (ACTFL) proficiency guidelines. At the test design stage, to ascertain the standard for a minimally competent test candidate on the CELBAN, test developers conducted an initial needs analysis to establish “the real-life English language demands of the nursing profession in Canada”, (p. iii) and to allocate suitable Canadian Language Benchmark (CLB) levels in the skill areas of Speaking, Listening, Reading and Writing (CELBAN, 2015). From this initial needs analysis, the Writing sub-test was linked to a CLB score of 7, being at Stage II, Intermediate Proficiency (Levels 5 to 8) (CELBAN, 2015). However, how this CLB linked score was actually decided on remains unclear as the passing cut score is not publically available.
In summary, as noted above, standard-setting studies conducted in the health arena have mostly involved general English proficiency tests, rather than LSP tests like the OET with tasks designed expressly to reflect workplace demands. The following section highlights recent standard-setting research conducted on the OET.

2.5.5.3 Standard-setting research on the OET

Apart from the Canadian English Language Benchmark Assessment for Nurses (CELBAN), the other known health-related, content specific language test is the OET. However, until recently, standard-setting studies on the OET have been limited and/or not available in the public domain (an exception is an early standard-setting study on the OET Speaking sub-test by Lumley, Lynch and McNamara, 1994). More recently, as part of a broader study on the OET Speaking sub-test, Pill and McNamara (2016) carried out standard-setting workshops with clinical educators and supervisors from three professions: nursing, medicine and physiotherapy. Their paper centred on two issues: the concept of clinical communicative ability and its separability (or not) from other characteristics of professional capability and the technical complexity of establishing a justifiable minimum standard of oral proficiency for overseas-trained health professionals to practice in an English-speaking workplace (a methodological issue). The study sought to answer the question of “how to determine what level of performance on the specific-purpose language test, scored by ESL trained raters, and as expressed through the test score, is sufficient in the eyes of health professionals to allow a health professional to participate safely and effectively in the Australian workplace” (Pill & McNamara, 2016, p. 2-3).

In the standard-setting workshops a modified 7-category system of the Analytical Judgement method (AJM) was successfully used by Pill and McNamara (2016), rather than the originally implemented 12-category approach of Plake and Hambleton (2001) referred to previously. Participants were recruited, in part, from an earlier phase of their larger project and totalled 39 overall: medicine – 13 panellists, 3 workshops, 26 samples graded; nursing – 18 panellists, 2 workshops, 25 samples graded;
physiotherapy – 8 panellists, 1 workshop, 24 samples graded. The authors concluded that, from a procedural and methodological perspective, the AJM was both practical and useful for conducting standard-setting workshops. The final cut scores varied somewhat across the three health professions and while the differences were only moderate, separate scores for each occupation were advocated. Pill and McNamara (2016) argue that while the reporting procedure associated with each level should not change (i.e. a grade of B should be maintained as the passing standard), profession-specific cut scores representing the different grade levels should be employed in operational administrations. The study asserted that permitting health professionals to establish cut scores produced more defensible conclusions about admittance to supervised clinical practice (Pill & McNamara, 2016). Two key questions were raised in their study: 1) the extent to which panellists followed the instruction to only judge test candidates’ language proficiency and not their professional skills/knowledge (as required by Australian government legislation); and 2) whether language issues could be satisfactorily separated from other clinical/professional skills on a language assessment because ‘subject-area specialists’ (in comparison to language specialists’) appreciation of what is important may be dissimilar or considerably wider in range (Pill & McNamara, 2016).

In answer to these above questions, a paper by Manias and McNamara (2016) further focused on the qualitative data gathered from the same standard-setting workshops described by Pill and McNamara (2016), for the purpose of judging the minimum standards of oral communicative proficiency for professional registration of overseas-qualified health professionals. Focus group commentary from the doctors and nurses participating in the standard-setting workshops was recorded and transcribed and a thematic analysis of the transcribed discussions was conducted “to determine the basis for the ratings given by participating health professionals to candidates’ role-play performances” (p. 5). The roleplay is a simulated clinical interaction where the test candidate assumes their professional role (e.g. as a nurse) while the interlocutor performs as a patient/client or sometimes a patient’s relative or carer. The main direction of Manias and McNamara’s (2016) findings was that in forming their judgements about standards, the HP domain experts were indeed, in the main,
attending to language related features of performance as they had been asked to do. However, judgements related to clinical competency from the OET clinical simulation were also made by panellists, even though they were explicitly instructed not to do so.

The analysis of domain experts’ views also attended to methodological issues that made it challenging to establish a minimum standard and centred on two main areas. The first was a “lack of clarity on the nature of the clinical context assumed, relevant to the standard”; or in other words, the potential problematic areas that domain expert standard setters might encounter in deciding on a minimum passing standard due to “the highly variable nature of workplace contexts, particularly in terms of the use of English as a lingua franca in the workplace both among health professionals and between health professionals and patients or clients” (e.g., in terms of the understanding of accents of test candidates by domain experts and use of idiomatic terms by test interlocuters) (Manias & McNamara, 2016, p. 16). The second challenge identified by Manias and McNamara (2016) in setting a minimum standard was “limitations of the roleplay” (p.20) due to authenticity concerns and its questionable link to everyday workplace practice. The participants raised questions of authenticity with the role-play tasks themselves (e.g., a typical task for medicine: “a 40-year-old single parent presents seeking advice and reassurance having experienced a ‘panic attack’ a couple of days previously” (p. 4)) such as whether the tasks were suitable to prove a test candidate’s ability to effectively communicate in the healthcare work environment (Manias & McNamara, 2016). In sum, the analysis of subject-matter experts’ views in this study yielded valuable insights into the standard-setting process and Manias and McNamara (2016) concluded that “such qualitative analysis should become a regular accompaniment of the quantitative analysis to which reports of standard setting are usually confined” (p. 24).

The next section examines several unresolved issues in standard-setting research. Some of these issues have yet to be addressed in detail in past standard-setting studies—this has particular implications for the present study as it attempts to answer some of these concerns.
2.5.6 Issues in standard-setting research

This section discusses three potential issues when conducting standard-setting research: variability, method effects and domain expert involvement.

2.5.6.1 Variability

Variability is one area of concern in standard-setting research. Kenyon and Römhild (2014) state that “standard settings are based on the judgements of multiple individuals, so variability in judgements is an expected outcome” (p. 5). Furthermore, variability is also highly likely between diverse groups with different professional backgrounds and training. They also note that too much variation in cut score results and issues with consistency of panellists’ judgements affects whether the test will continue to be used (Kenyon & Römhild, 2014). Similarly, Brown (2013) argues “decisions made on cut scores are error prone, especially near any given cut score” (p. 3). Therefore, attention to the sources of variation in standard-setting judgements and their potential impact on the stability of cut scores is critical in any standard setting research. This study takes into account and attends to possible variability issues with participants’ judgements (see method section 3.6.4.5.2).

2.5.6.2 Method effects

Another potentially problematic issue in standard-setting studies is method effects. Kane (1994) states that validating the results of any one procedure can be challenging and complex, and as mentioned, there may be threats to a chosen method’s validity. These might include inconsistency in the application of a method, misunderstandings about what is entailed in the procedure from inexperienced panellists, and fatigue interfering with standard-setting judgements due to the time-consuming nature of a method. Kenyon and Römhild (2014) note that to ensure standard-setting practices are valid and justifiable some crucial features require careful attention such as “the selection and training of panels of judges, the incorporation of consensus-building
processes and feedback information, and the collection and documentation of evidence on the procedural validity of the standard setting” (p. 3-4).

From the previously above-mentioned health profession studies, it can be seen that multiple standard-setting methods have been used with general English language proficiency tests and the fact that a variety of methods are in use (and sometimes with inconsistent application) makes it difficult to generalize from one study to another. Cizek (2012) states that in early research into standard setting there was relative homogeneity of implementation of methods and some studies compared two or three methods with their relative performance described. In a language testing context, a recent example of this kind of comparative approach is a study by Shin and Lidster (2016) that investigated three standard-setting methods for ESL placement testing at a large Midwestern university in the United States. Their study compared the strengths and weaknesses of the Bookmark method (a test-centered approach, described earlier), the Borderline Group method (an examinee-centered approach, described previously) and hierarchical cluster analysis (a statistical approach). The study found that different methods yielded different cut score outcomes and, in light of the findings, evaluated the benefits and limitations of each method with implications for future usage. Due to the evident disparity between methods and participants’ ratings, the study recommended more research on “the actual decision-making processes of raters” (Shin & Lidster, 2016, p.22). The current study attempts to address this decision-making issue.

With the recent proliferation of adaptations and modifications of traditional methods and their implementation as described in the literature, a lack of agreement in reaching cut scores between methods is not surprising because at present Cizek (2012) states “the degree of heterogeneity among applications of any one method is remarkable” (p. 11). He laments that this may signal “the end of method” as “the characteristics of standard-setting procedures for a specific application vary from the characteristics of the same procedure applied in other contexts, it becomes increasingly difficult to conduct focused research on a method, or even to operationalize the distinctive characteristics of the method one hopes to study” (p.
11). He rather despondently points out that “increasingly, standard setting may be evolving as a ‘one-off’ technology” (Cizek, 2012, p. 11). Because of the high associated expense of running standard-setting workshops, and as noted for reasons of participant recruitment and subsequent generalizability of results, online panels are worth considering. Recent technology-based directions, as noted by Harvey and Way (1999), Katz, Tannenbaum and Kannan (2009) and Katz and Tannenbaum (2014), discussed the feasibility of using standard setting methods online and the web-based setting of cut scores. With increasing levels of technological use and attention in many aspects of language testing and assessment, Katz and Tannenbaum (2014) suggest though, as in all standard-setting workshops, “technique trumps technology” (p. 16).

Considerable research (as noted above) has concentrated on mainstream standard-setting methodologies and associated variations aimed at establishing valid cut scores between performance levels on a particular test. Standard setting has to date had a predominantly quantitative focus. Buckendahl (2005) argues “the challenge for all standard-setting methodologies is to effectively translate a participant’s mental model of the target examinee (e.g., barely proficient student) into judgements that communicate the participant’s recommendation of a value that characterizes the point of separation between one or more categories” (p. 219). Most studies thus far have centred on the procedures involved in arriving at the actual ‘point of separation’ or cut score, but relatively little attention has been given to understanding participants’ conceptual paradigm or the nature of thought processes underlying standard-setting judgements.

Some recent exceptions that have had a qualitative focus include Hein and Skaggs (2009) who conducted post standard-setting workshop focus group interviews to investigate participants’ feelings and experiences of the process. Hein and Skaggs (2010) also examined the cognitive processes of a group of teachers in conceptualizing a theoretical group of target students by again conducting post workshop focus group interviews. Other studies such as Giraud and Impara (2005), Giraud, Impara, and Plake (2005), Ferdous and Plake (2005), McGinty (2005) and Skorupski and Hambleton (2005) undertook qualitative research that included aspects such as assessing
panellists’ definition of a target test candidate, asking participants to write down their thoughts at various points throughout the standard-setting process, conducting interviews after the standard-setting task in order to determine features that impacted panellists’ judgements, noting the interactions between participants and facilitators in standard-setting sessions and observing a school region’s standard-setting process and the experience of stakeholders in the procedure. The data and comments from panellists in these qualitative-focussed standard-setting studies (as is the case in the majority of standard-setting studies that have a qualitative element) were used as part of the process of confirmation of the procedural validity of the standard-setting method. Claims about the validity of any standard-setting workshop rely on, at least partly, the qualitative feedback of participants on the training and procedure overall, usually in the form of a final evaluation. However, Kenyon and Römhild (2014) state that qualitative data often collected from panellists during standard-setting studies to justify procedural validity, “tends to be superficially analysed and little research is actually done on the cognitive and social processes at work in the standard-setting process” (p. 14). Buckendahl (2005) also argues while there have been theories put forward on the cognitive requirements for standard-setting panellists, few studies have investigated judges’ thought processes or feelings while being a participant in judgmental standard-setting procedures. Kenyon and Römhild (2014) advocate the use of unique qualitative approaches to further clarify how panellists’ judgements come about and to explore issues related to construct validity. It is also suggested that more discourse analyses of panel discussions that seek to build consensus in cut score judgments are needed (Kenyon & Römhild, 2014). Close analysis of these panel discussions may give insight into how decisions are made and, ultimately, serve to enhance discussion group procedures.

In a language-testing context, standard-setting studies that have a qualitative focus are scarce. In a study that considered the judgement-making process of participants in aligning a language test to the Common European Framework of Reference (CEFR), Papageorgiou (2010) showed the value of using qualitative data collection techniques in standard-setting research. The study analysed the focus group discussions of panellists between rounds of judgements and discovered that participants’ decision-
making was influenced by aspects that were not relevant to the judgement task. The thematically coded qualitative data showed that deciding on CEFR cut scores by participants was at times problematic and illustrated some of the inherent issues with the CEFR such as the wording of its descriptors and context-free nature of its scales. The study did though “demonstrate the usefulness of qualitative analysis in obtaining a better understanding of the judge’s perspective in the CEFR cut score setting context” (p. 276). However, the panellists used in the study were not domain experts, but rather language experts and the focus was on general language proficiency. Also, the paper by Manias and McNamara (2016) (described previously) has contributed to the literature, but more of this type of qualitative research is necessary.

2.5.6.3 Domain expert involvement

The inclusion of domain experts in standard setting is somewhat taken for granted and researchers have not thus far questioned the validity of domain expert judgements to any great extent. There tends to be an assumption that the views of panellists setting the standard correspond with criteria used to assess performance (i.e. they are drawn from same underlying construct) (Kenyon & Römhild, 2014). However, the degree to which this is true is uncertain. Whether domain experts are well equipped to establish valid standards on an LSP test such as the OET is unclear and warrants further investigation through qualitative research into their decision-making practices.

The only recent qualitative research that has explored subject-matter experts’ judgement-making in standard setting for healthcare communication is, as previously mentioned, Manias and McNamara’s (2016) published paper. Their qualitative study analysed the workshop discussions between standard-setting participants and focussed on the foundations of how judgements were made. A necessary component of the Analytic Judgement method (AJM) (as used in this study and in many standard-setting methods) is for participants to discuss and compare their performance level allocations after a round of judgements. Manias and McNamara (2016) argue that even though subject-matter experts may be involved in LSP standard setting, in the
actual test arena it is usually language experts instead of domain experts who rate the
tests and that “there is therefore a possible tension between what is being measured
by a test and what is actually deemed important in the domain” (p. 2). As previously
noted, the insights of domain experts in standard setting are valuable as they may
bring to light issues that would not be attended to by language testers. In a health
communication environment this could be related to important concerns such as the
safety and wellbeing of patients involved. Also, involving subject-matter experts in
standard-setting procedures may lead to conclusions being drawn from their insights
about the boundaries of what a language test is able to gauge. In addition, domain
expert involvement may help to confirm that the test is measuring what it is intended
to measure instead of other non-relevant aspects. Likewise, subject-matter expert
involvement might point to features that are not wholly construct relevant, yet are
considered important to them. For example, they might value or give more weight to
certain attributes which, while important in their professional context, may be less
central to judging performance in the necessarily artificial environment of the test (see
discussion section 5.3.2.2 on Patient Identification).

Further qualitative research may shed understanding of whether standard-setting
participants draw on relevant constructs while making judgements and also if the
judgements made are consistent between judges and individual judgements. If there
is inconsistency between participants, possible reasons could be due to their lack of
familiarity with the test construct or even their capability to carry out the task. There
is an assumption that domain experts have the relevant insights to make appropriate
standard-setting judgements over non-subject-matter experts, yet this may not
always be the case. Kenyon and Römheld (2014) urge that further qualitative research
is necessary in standard-setting studies to investigate these complex issues. Pill and
McNamara (2016) also argue that in order to further validate standard-setting
procedures and explore possible variation between participants’ judgements
“research, for example using verbal protocols to study the raters’ decision-making
processes, is called for” (p.15). In response to this call, the current study attempts a
more in-depth analysis of what informs domain experts’ judgements when setting
standards on the OET Writing sub-test.
The previous sections reviewed studies related to language testing and standard setting and health profession-related studies. Standard-setting research on the OET and the general lack of such research in the field of language testing was highlighted. The involvement of domain experts or not in a standard-setting study was considered. Current concerns in standard-setting research were outlined including the need for qualitative standard-setting studies which illuminate the nature of decision-making behaviour. Few studies offer insight into what standard-setting judges bring to bear in setting standards. The following section examines test validity concerns and in particular a framework for evaluating validity that may be used for an LSP test such as the OET.

2.6 Test validity

This section reports on the notion of test validity. First, the concept of validity and historical changes in thinking about this concept are discussed. Second, current ‘argument-based’ approaches to validation are reviewed and the strengths and weaknesses are noted. Finally, the direction that the current study takes in regard to validation is outlined.

2.6.1 Historical changes in conceptualizations of validity

Validity has been explained in a number of traditions and the ways it is defined and conceptualized have changed over time. Chapelle (2012) mentions the “evolving or shifting concepts of validation” (p. 119) put forward by Anastasi (1986), Angoff (1988), Kane (1992, 2001, 2012a), Messick (1989), and Moss (1992). A number of sometimes conflicting and other times converging views underpin these theories. This is significant as the manner in which concepts of validity are formed may shape the character and range of a validation study and therefore the means employed to collect evidence. These shifting formulations are elaborated in this section.

Several approaches and conceptualizations of validity have been discussed over time (Chapelle, 1999, 2012; Xi 2008). Early theories of validity in educational measurement
talked about different ‘types’ of validity (e.g., ‘criterion’, ‘content’ and ‘construct’ validity). Thorndike (1918) focused on the ‘criterion’ measure that the test was designed to predict, exploring the association between test scores and criterion scores (e.g., training grades, job knowledge test scores and/or supervisor ratings). A robust association between the test scores and relevant criterion measures was taken as evidence of the test’s predictive or criterion validity. Until the 1950’s, Thorndike’s approach was thought to be the ‘gold standard’ for validity studies (Angoff, 1988; Cronbach, 1971; Moss, 1992). ‘Content’-based validation methods (i.e. demonstrating that test items replicated the knowledge actually needed for a particular subject area) were also used from the 1950s for interpretations that centred on well-delineated performance contexts such as tests of specific skills (e.g., accounting), but did not work as well for interpretations outside of such contexts (Kane, 2004). Cronbach and Meehl (1955) put forward ‘construct’ as another possible focus for test validation studies. To claim construct validity for a test, it was necessary to demonstrate via statistical analysis that test scores quantified the proposed construct. Construct validity is related to the main ‘trait’ or construct being assessed by the test and remains a central focus of test validation work (McNamara & Roever, 2006). How the construct is defined will directly influence the meaning of the test scores (Hubley & Zumbo, 2011).

From the 1960s, and in regard to language testing, validity was seen as an essential ‘quality’ of a language test and it was mainly thought to be associated with the test itself and test scores (Bachman, 1990; Chapelle, 1999; Kane, 2001; Messick, 1989). The writings of Lado (1961), Heaton (1975) and other language testing scholars (e.g., Harris, 1969; Valette, 1977) in the field supported this view. In the late 1980s/early 1990s, a re-examination of the definition of validity uncovered its complex and multifaceted nature. Messick (1989), working in the field of educational measurement, argued that validity is not a quality of tests themselves, but rather an integrated evaluative decision concerning whether empirical confirmation and theoretical foundations can support the suitability of conclusions founded on assessment scores. In his view, validity is a unitary concept with construct validity incorporating all other facets. Anticipated outcomes of an assessment should be mapped out at the planning phase and data subsequently collected to establish if
results match up with what was envisioned (Messick, 1989). This ‘accumulation-of-evidence’ or ‘evidence-gathering’ approach (Chapelle, 1999; Kane, 2001) uses a set of substantiations to back or argue against a particular test score interpretation or use. More recent language assessment research (Bachman, 1990; Bachman & Palmer, 1996, 2010; Chapelle, 1999, 2007; Weir, 2005) built on Messick’s (1989) conception of validity and a number of authors introduced the role of ‘interpretive arguments’ (Kane, 1992, 1994, 2001, 2004, 2006, 2012a, 2012b, 2013; Kane, et al., 1999a; Kane, Crooks & Cohen, 1999b) to ensure that test score uses and interpretations were clearly defined and validation was more workable for assessment developers and researchers. The next section examines one such method, the ‘argument-based’ approach.

2.6.2 Argument-based approach to validity

The ‘argument-based’ approach to test validation has undergone much refinement and development. Toulmin’s (2003) influential argument model involves making claims founded on data from the test context. Toulmin’s structure has guided the development of a number of recent validity frameworks (e.g., see Bachman & Palmer, 2010; Chapelle, 2011; Kane, 2006, 2012a, 2012b, 2013). These arguments have informed the revision and validation of language tests (Chapelle, Enright, & Jamieson, 2010) or the development of validity frameworks in language testing (Bachman, 2005; Bachman & Palmer, 2010). An argument-based approach (Cronbach, 1988; House, 1980; Kane, 1992, 2006; Shepard, 1993) to test validation, charts a method of test use and development that justifies the assessment procedure. Kane (2013) notes an argument-based framework consists of firstly stating comprehensively the intended interpretations and uses of test scores and secondly appraising the credibility of the intended interpretations and uses taken as a whole. This means that when applying an argument-based approach, an ‘interpretive’ argument is put forward which is based on a number of inferences and bolstering assumptions that link scores to conclusions and judgements (Bachman, 2004; Crooks, Kane & Cohen, 1996; Kane et al., 1999a, 1999b; Kane, 2002). A ‘validity’ argument gives an overall appraisal of the plausibility of the interpretive argument. Kane (1992) states that the interpretation of
Test scores need to be founded on an argument that is thought to be reasonable and the plausibility of an interpretive argument requires several inferences and assumptions. The fundamental basis of an interpretive argument is founded on the basic constituents of ‘test performance’, ‘test score’, ‘universe score’ and ‘test taker’s ability in real life’ (Kane, 1992). These are bridged by broad inferences in test score interpretations: ‘evaluation’, ‘generalization’, ‘explanation’ and ‘extrapolation’, ‘decisions’ and ‘test consequences’, that all need supporting evidence (Crooks et al., 1996, Kane, 1992). To correctly make sense of and comprehend a test score, the inferences and meaning of the score must be made clear. Figure 2.1 below represents the key constituents of an interpretive argument.

Figure 2.1: Building blocks of an interpretive argument (adapted from Kane (1992) in Knoch & Elder (2013)).

In this figure, each arrow signifies an inference which acts as a bridge between one part of the argument and the next. Each bridge requires supporting evidence for a validity argument to be upheld and the final interpretation to be considered as valid. In addition, each inference has a range of underlying assumptions that need to be clearly put forward to serve as a foundation for evidential research support (Kane, 1992; Knoch & Elder, 2013).

The inference linking ‘test performance’ with ‘test score’ is the ‘evaluation’ inference and comprises converting a test candidate’s assessment performance into a defensible test score. Underlying assumptions which need evidential support are that the test’s scoring procedure is performed reliably and that the scoring system truly mirrors the test construct or what the test is intended to assess. Furthermore, there is an assumption that the test instructions or the test administration procedures have not presented any construct-irrelevant variance or aspects that could impede what
the test is intended to measure. Whether or not these assumptions can be supported will allow the raw score to be regarded as defensible. The next inference bridging a candidate’s ‘test score’ with the ‘universe score’ is termed ‘generalization’ which assumes that irrespective of specific tasks or judges, the test score resulting from any test occasion is demonstrative of the score a test candidate would be given on other test versions or administrations (i.e. commonly referred to as test ‘reliability’). Lastly, the ‘explanation’ and ‘extrapolation’ inferences link the ‘universe score’ with the ‘target score’ (i.e. the connection between the theoretical construct that the assessment is intended to measure and as a consequence a test candidate’s performance in a real-life context). An underlying assumption necessitating evidential support would comprise how well test tasks draw out from test candidates the assessment designers’ intended target skills – as inferred in the construct of ability on which the assessment is based (i.e. the ‘explanation’ inference). Another assumption would be how well the test tasks replicate the language requirements of the applicable real-world domain (i.e. the ‘extrapolation’ inference) (Kane, 1992; Knoch & Elder, 2013).

Later adaptations also included ideas of ‘test consequences’ or ‘test use’ (Kane, 2001, 2002, 2004, 2006, 2012a, 2012b, 2013; Kane, et al.,1999a, 1999b) which took into account an additional constituent, ‘decisions’, and the inference ‘test use/consequences’. The decisions and test use inferences have particular relevance to standard setting. In many instances, assessments are explicitly intended and designed to back particular decisions about test candidates such as selection, placement and diagnosis by delivering information about test-candidate characteristics that are pertinent to the decision (Kane, 2013). In addition, assessments are developed to evaluate aspects which may be applicable to a variety of decisions in a number of contexts (e.g., a test of communicative competence in a particular language might point to a test candidate’s ability to utilize that language successfully in a range of contexts). Subsequently, the assessment may then be employed by stakeholders to make a number of decisions about test candidates (Chapelle, 1999).
In addition to inferences and underlying assumptions, the formation of a validity argument involves claims built on the basis of data and warrants (Bachman, 2005; Toulmin, 2003). In this account, a claim is a conclusion whose worth is ascertained, or in other words a claim is an interpretation formed on the basis of data about what a test candidate can do or knows. Data includes an account of what the claim is founded on and in the case of language testing, these are the responses of test candidates to assessment tasks. Warrants and rebuttals serve as a linkage between data and a claim and need to be thoroughly scrutinized in relation to their construction and nature (Toulmin, 2003). A warrant is a proposition that is used to validate an inference and is characterized as a broad statement that offers authenticity of a specific stage in the argument. In addition, the backing of a warrant is based on other assertions without which the warrants themselves would not hold confidence or prevalence. On the other hand, a rebuttal puts forward alternative explanations or counterclaims to the proposed inference and is comprised of extraordinary circumstances which may be able to beat or refute the conclusion. Rebuttal data contains verification that might corroborate, weaken or reject counterclaims (Toulmin, 2003).

In an argument-based approach, an interpretive argument methodically presents a series of inferences and claims connected with test score explanations and subsequently the credibility of these inferences and claims are assessed employing empirical and theoretical support (Kane, 1992). Validation involves an unambiguous account of the planned interpretations or uses and claims and inferences of test scores (Kane, 2006). Furthermore, an argument-based approach considers test validation as a continual and critical process to develop a validity argument for a specific examination (Chapelle, 1999; Kane, 1992, 2001). Kane’s (2006) interpretive argument realization necessitates a reflective procedure that entails the test’s validity argument be built on theoretical rationales and empirical data to support score significance and use. This direct evidence-based approach to argument-based validation and the use of interpretive arguments can be evidenced by Chapelle (2011) and Chapelle, Enright and Jamieson (2010) in their framing of a validity argument for the Test of English as a Foreign Language (TOEFL). This approach was also used successfully by Knoch and Elder (2013) when they presented a range of arguments to support validity claims in
post-entry English language assessment (PELA) contexts and demonstrated the kinds of evidence needed to support these validity arguments.

Kane (1992, 2002) mentions that dissimilar to mathematical or logical arguments, assumptions in an interpretive argument cannot be presumed as certain. This is because the evidence supporting these assumptions is sometimes debatable or incomplete. Furthermore, McNamara and Roever (2006) assert “validity is not a mathematical property like discrimination or reliability, but a matter of judgement” (p. 10) and shows the subjectivity of judgements formed in connection to the validity of a test. The conclusions of interpretive arguments are therefore not confirmed – they can only be assessed in terms of how plausible or believable they are. Kane (1992, 2002) notes three criteria for appraising the inferences of an interpretive argument: 1) clarity of argumentation (i.e. the argument should be asserted plainly so that what it claims and presumes are known); 2) coherence of argument (i.e. the conclusions should be reasonable and logical given the assumptions); and 3) plausibility of assumptions (i.e. the assumptions need to be believable or confirmed by evidence that could comprise parallel lines of evidence or credible refutations/counterarguments).

Bachman and Palmer (2010) in their conceptualization of validity argued that Kane’s ‘interpretive’ argument term may have given too much emphasis to test ‘interpretations’ and not enough to test ‘uses’ and therefore proposed a ‘test usefulness’ approach that has six characteristics: ‘reliability’, ‘construct validity’, ‘authenticity’, ‘interactiveness’, ‘impact’ and ‘practicality’. This methodology is meant to ensure that the tasks in a language test are useful for their planned intentions and relate in a verifiable procedure to language use tasks (Bachman, 2001). This ‘usefulness’ model has been further elaborated and arguably superseded by an ‘assessment-use argument’ (AUA) framework (Bachman & Palmer, 2010) which first formulates specific claims that connect test takers’ performance to the effects of test usage and second gathers evidence to back up the stated claims. This “extended what Kane proposed in the area of decisions and test consequences, offering a range of use and consequence-related assumptions that need to be supported for the score-based interpretations and uses to be valid” (Knoch & Elder, 2013, p.52). Bachman and
Palmer’s (2010) AUA approach has been characterized as an ‘assessment justification’ that “focuses on investigating the extent to which the intended uses of a particular test can be justified to stakeholders” (Wang, Choi, Schmidgall & Bachman, 2012, p. 603). The next section highlights the strengths and weaknesses of these two key frameworks for evaluating language test validity for their possible use in framing the discussion of this study.

2.6.3 Strengths and weaknesses of an argument-based approach

The two frameworks of validation, the ‘argument-based approach’ proposed by Kane (1992) and the ‘assessment use argument’ (AUA) offered by Bachman and Palmer (2010) have their own strengths and weaknesses. One of the strong points of the argument-based approach (also true of an AUA) is the support it offers in assigning actual research directions and indicating the specific type of validity evidence that is required (Bachman, 2004; Cronbach, 1988; Kane, 1992). The construction of the interpretive argument establishes the type of evidence that needs to be collected at each stage of the validation process and affords a foundation for assessing progress as a whole. The approach does not state that one type of validity evidence is better than any other, however the choice of evidence should deal with the credibility of the particular interpretive argument being offered (Kane, 1992). The sorts of validity evidence that are likely to be most appropriate would pay particular attention to the inferences and assumptions that are most open to challenge. An interpretive argument’s most unconvincing elements should be the focal point of the evaluation and the interpretive argument may have to be amended or discarded if some inferences are discovered to be untenable.

An interpretive argument does not bring about an outright judgement concerning validity. However, it does offer an approach to measuring progress (Kane, 1992). A benefit of this method is that it considers test validation as a critical and continuing procedure rather than a static one with a definite decision of ‘valid’ or ‘invalid’. In this process, the inferences that are most uncertain can be re-evaluated with further evidence or they can be fine-tuned to make them more credible. Consequently, the
overall credibility of the interpretive argument can be strengthened. The rigour of this approach is evidenced by its acknowledgment that an interpretive argument’s reader is one who requires convincing. Also, there is a necessity to advance a clear case for the suggested inference, and there is a requirement to examine and assess possible opposing interpretations. By analyzing test validation in such a way, an understanding of the key stages of assessment development can be appreciated.

Nevertheless, uncertainty may still exist about how to employ an argument-based approach in practice. Bachman (2004) contested that the interpretive argument-based perspective on validation in language testing (Kane, 1992, 2002) did not thus far adequately tackle the concern of test use and its consequences as a facet of test validity and that this issue should be attended to in further validation research that employs an interpretive argument. In response to this lack of score uses and consequences in the defining of an interpretive argument, Kane (2013) updated the term to now be delineated as an ‘interpretation/use argument’ (IUA). This was done “in order to recognize the importance of score uses in determining score interpretations and to acknowledge the importance of score uses (as well as contexts and test-taker populations) in validation” (Kane, 2013, p. 65).

According to Tannenbaum and Cho (2014), the assessment use argument (AUA) framework conceptualized by Bachman (2005) and Bachman and Palmer (2010) is a valuable one as it further develops the earlier foundations of argument-based approaches to validation such as Kane (2006, 2013) by incorporating measures for examining the usefulness of assessment scores and the anticipated results as a consequence of utilizing such assessments (which is of interest in standard setting). Kane (2013) also later acknowledged score uses as essential in validation (as noted above) and has now accommodated Bachman’s objections by adding test decisions and consequences. Therefore, there is nothing in particular to recommend one framework over the other. However, as Kane’s (1992) framework has been highly influential and has served as a foundation, this study utilizes his argument-based approach (with the later insertion and acknowledgment of test decisions and test consequences (i.e. Kane, 2013) drawn from Bachman and Palmer’s (2010) assessment use argument (AUA) framework). Decisions and test consequences have particular
importance for standard-setting studies (as explained below). The use of Kane’s (1992, 2013) argument-based approach in application to LSP testing has so far been under-utilized in the literature and this study contributes to the knowledge base of validation studies using this framework. The following section examines how standard setting fits into an argument-based approach to validity.

2.6.4 Standard setting in an argument-based approach

As demonstrated above, there is a considerable body of literature on standard setting in the field of educational measurement, and more recently, in language testing. However, Papageorgiou and Tannenbaum (2016) point out that the place of standard setting within argument-based approaches to test validation has not always been well-defined; also noted by Bejar, Braun and Tannenbaum (2007), Kane (2004), McClarty, Way, Porter, Beimers and Miles (2013), and Pant, Rupp, Tiffin-Richards and Köller (2009). A similar point has been made by Kenyon and Römhild (2014) in relation to standard-setting research in the language testing field.

As mentioned, in an argument-based validity framework, evidence needs to be collected to support claims made about test scores’ uses/consequences. In many standard-setting studies, as seen in the educational measurement and language testing literature (see e.g., Council of Europe, 2011; Hambleton & Pitoniak, 2006; Kaftandjieva, 2004; Kane, 1994; Tannenbaum & Katz, 2013) particular cut score validity criteria are usually employed to assess the evidence. The type of validity evidence in standard-setting studies typically deals with three main ‘types’: ‘procedural’, ‘internal’ and ‘external’ validity. Procedural validity is concerned with the authority or reliability of the procedure itself and “whether the procedures followed were practical, implemented properly, whether feedback given to the judges was effective, and whether documentation has been sufficiently compiled” (Papageorgiou & Tannenbaum, 2016, p. 111); internal validity involves the consistency and accuracy of the standard-setting results; and external validity requires “evidence from independent sources that support the outcome of the standard-setting workshop” (Papageorgiou & Tannenbaum, 2016, p. 111). While support for these categories of
validity has been commonly sought in many earlier standard-setting studies, it is perhaps more appropriate, in keeping with current thinking about validity, to talk of assumptions underlying distinct validity inferences, rather than to depict validity as being of different ‘types’. This characterization of validity as being of different kinds is somewhat outdated and the assumptions which they support have not been explicitly formulated in the context of an argument-based validity framework. The framing of validation within an argument-based approach is quite dissimilar to the three types (i.e. procedural, internal and external validity) mentioned previously, as is shown below.

In an argument-based approach, Kane (1994) asserts that the validity of test-based judgements concerning a test candidate’s preparedness or not hinges on the suitability of the passing scores employed to generate the judgement. He states cut scores need to be substantiated with “judgemental standard-setting studies” (1992, p. 12) and notes high-stakes tests such as certification tests (like the OET) are developed to defend decisions about “safe and/or effective performance in practice” (p. 161). A ‘warrant’ is a directive stating how the decision will be formulated, and therefore, according to Kane (1992): “The cut score corresponds to the performance standard, in the sense that persons with scores above the cut score have generally achieved the performance standard, and that persons with scores below the cut score have generally not achieved the performance standard. The backing for the warrant needs to support these two assumptions” (p. 12).

The backing for the decision directive is intended to substantiate that the decision method will accomplish an objective (e.g., public safety) at a tolerable price (i.e. positive costs should prevail over negative ones) (Kane, 2004). The selection of a passing score is an essential concern and the backing for this score is established on value judgements about “how good is enough” (Kane, 2004, p. 161). Cizek (2001) notes that empirical evidence from a variety of stakeholders should be used to offer collective judgement and acceptable backing for the passing score. The fundamental issue is about what is the degree of test performance required that mirrors the degree of competence in the real-world domain (Kane, 2004). Therefore, standard-setting
studies must substantiate the grouping of test takers as above or below a cut score and as a result represent their use as credible (Oller, 2012).

Kane (1994) argues the interpretation of passing scores is governed by two suppositions: 1) the passing score matches the specific performance standard (i.e. test candidates with scores above the passing score are liable to achieve the standard and those below are not); and 2) the specific performance standard is realistic in terms of the aim of the decision. These two suppositions can be assessed on the correspondence between the methods employed to establish the passing scores and the intention of the decision, the internal reliability of the outcomes, and contrasts with external criteria. The basis of inaccuracies in the passing score can be distinguished by analyzing these two suppositions (Kane, 1994).

As stated, an argument-based approach to validation comprises a set of inferences and assumptions from test performance to test consequences (Kane, 2004). The ‘decisions’ inference and the setting of standards and cut scores on the OET Writing sub-test are of particular interest to this study. Decisions presumes that the setting of minimum standards for professional registration that are founded on the OET Writing sub-test are suitable and fair. In high-stakes LSP testing, not involving domain experts in the decision-making procedure of setting passing standards and cut scores may influence the validity of the cut scores and might cause doubt about the connection between the test use and consequences for stakeholders.

Papageorgiou and Tannenbaum (2016) utilized Bachman and Palmer’s (2010) ‘Assessment Use Argument’ (AUA) framework to demonstrate how evidence from standard setting can sustain claims about consequences, decisions and interpretations. They argue for the utilization of an AUA framework, as opposed to any other, “because of its emphasis on the use of language assessments, in particular the decisions that should be made on the basis of the use of these assessments and the consequences of the decisions for students, teachers, and the greater educational and social context” (Papageorgiou & Tannenbaum, 2016, p. 113). The four claims adapted from Bachman and Palmer (1996, 2010) in which evidence would be needed
to support/refute these claims were: 1) the consequences of using an assessment and of the decisions that are made are beneficial to all involved stakeholders; 2) decisions that are made on the basis of the assessment-based interpretations take into consideration community values and relevant legal requirements and are equitable for those stakeholders who are affected by the decisions; 3) interpretations about the ability to be assessed are meaningful with respect to a syllabus, theory or TLU domain, impartial to all groups of test takers, generalizable to the TLU domain, relevant to the decision to be made and sufficient for the decision to be made; 4) assessment records (e.g., scores, descriptions) are consistent across different assessment tasks, different aspects of the assessment procedure and across different groups of test takers (Papageorgiou & Tannenbaum, 2016, p. 114). Their study attempted to set cut scores for the Common European Framework of Reference’s (CEFR) (Council of Europe, 2011) six main proficiency levels and related classification scales/descriptors.

Even though they concluded that standard setting should not be treated as an isolated event in the development of tests and should be essential to a validity argument, a number of issues were noted by the authors themselves “including challenges in relation to building an argument for test use and the limitations of the CEFR as an instrument for setting cut scores” (p. 119). Contrary to usual standard-setting panels (as utilized in LSP standard-setting studies) where the participants co-construct and define ‘Performance Level Descriptors’ (PLDs) themselves, their study used prior-defined CEFR descriptors to function as PLDs at a particular level. They found that “consequently, a challenge with the use of such a priori defined PLDs is that their description of language skills and abilities might not be directly relevant to the performance demonstrated when taking the test, as typically expected when panellists develop their own PLDs” and therefore “because panellists refer to the PLDs to make their cut score judgments, the CEFR descriptors might be difficult to use as such” (Papageorgiou & Tannenbaum, 2016, p. 118). Another drawback of the AUA framework employed in Papageorgiou and Tannenbaum’s (2016) study, in connection to the CEFR (a general language proficiency test), is that it did not address some of the issues related to standard setting in an LSP setting, where particular kinds of evidence are required to validate claims about the suitability and relevance of the test to the
real-world occupational setting. Hence, a more suitable frame needs to be utilized to examine an LSP test such as the OET and its standards.

Fortunately, in a recent paper, Knoch and Macqueen (in preparation) have developed an overall validity argument for LSP tests in general. The framework lays out an interpretive argument (Kane, 2004) by posing specific claims, warrants and assumptions applicable to LSP testing situations and listing potential sources of backing (evidence) required to confirm/refute these claims, warrants and assumptions. The inferences and associated claims they mention that require supporting evidence/backing include: ‘domain description’, ‘evaluation’, ‘generalization’, ‘explanation’, ‘extrapolation’, ‘decisions’ and ‘test consequences/ramifications’. The claims associated with standard setting are grouped under the ‘decisions’ inference, and are of particular interest to the current study. Knoch and Macqueen’s (in preparation) argument-based validity framework for an LSP test is a unique and useful one that could be applied to the present study on the OET Writing sub-test.

The ‘decisions’ inference, as formulated in Knoch and Macqueen’s (in preparation) overall validity argument for an LSP test is particularly relevant to the focus on standard setting in the current study. The inference “assumes that the decisions made based on the LSP test are appropriate and equitable” and is accompanied by the claim that such decisions “based on the estimates of the quality of the performance are appropriate and well communicated” (Knoch & Macqueen, in preparation). When applied to standard setting and the OET Writing sub-test, the warrant that “estimates resulting from the performance on the LSP tasks are useful for decision-making about readiness (or similar purposes) for work in the TLU domain” (column 1 in the table below) and related assumptions (column 2) are linked to potential sources of backing (column 3). Actual evidence/backing gathered for the present study is noted in method section 3.7. The findings from the current study are summarized around Knoch and Macqueen’s (in preparation) argument-based validity framework for an LSP test (such as the OET Writing sub-test) in conclusion section 6.2.
**Decisions inference**: assumes that the decisions made based on the LSP test are appropriate and equitable

**Claim**: decisions made based on the estimates of the quality of the performance are appropriate and well communicated

<table>
<thead>
<tr>
<th>Warrants</th>
<th>Assumptions</th>
<th>Potential sources for backing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimates resulting from the performance on the LSP tasks are useful for decision-making about readiness (or similar purposes) for work in the TLU domain</td>
<td>The standards set on the LSP test reflect the language standards operating in the TLU domain</td>
<td>Involvement of domain experts in standard setting; interviews with domain insiders receiving successful test takers</td>
</tr>
<tr>
<td></td>
<td>Standard-setting panellists are oriented to construct-relevant features of the performance</td>
<td>Verbal protocols during standard-setting; discussions during workshops</td>
</tr>
<tr>
<td></td>
<td>Variation in qualitative and quantitative judgements of standard-setting panellists is within expectations (or can otherwise be controlled)</td>
<td>Statistical analysis of score data; analysis of verbal protocol/discussion data</td>
</tr>
<tr>
<td></td>
<td>There is sufficient and appropriate information for the standard-setting panellists to make a decision</td>
<td>Interviews, surveys with standard-setting panellists</td>
</tr>
<tr>
<td></td>
<td>Standard-setting panellists are confident in the validity of their judgements</td>
<td>Feedback from standard-setting panellists</td>
</tr>
<tr>
<td></td>
<td>Standard-setting procedures are suitable and consistently applied</td>
<td>Justification for choice of method; detailed account of procedures; feedback from standard-setting panellists; comparisons with other standard-setting studies using the same method</td>
</tr>
</tbody>
</table>

**Table 2.1**: Decisions inference based on an LSP test (Knoch & Macqueen, in preparation)

Each of the assumptions listed in the above table requires evidential backing of the kind formulated in Column 3. Firstly, the assumption that “standards set on the LSP test reflect the language standards operating in the TLU domain” can be backed up with evidence from domain experts from the relevant occupational setting. For
example, domain experts could be interviewed about their experiences of working with successful test takers classified by the test as being ready to communicate in the particular workplace context. Secondly, the assumption that “standard-setting panellists are oriented to construct-relevant features of the performance” needs to be confirmed or rebutted – the implication being that domain experts involved in standard setting will be making decisions about workplace readiness based on factors that are relevant to the ability being tested. This can be established by probing the basis for their decisions by analyzing the discussions that take place during standard-setting workshops (as was done by Manias and McNamara (2016) in relation to the OET – see above) or through targeted elicitation procedures such as verbal protocols. Thirdly, the assumption that “variation in qualitative and quantitative judgements of standard-setting panellists is within expectations (or can otherwise be controlled)” can be established by analyzing the statistical data from standard-setting workshops to ensure that there is sufficient consistency in judgements to yield a robust outcome, and again, by analyzing discussion data or verbal protocol data associated with these judgements to determine any sources of variability. Next, the assumption that “there is sufficient and appropriate information for the standard-setting panellists to make a decision” can be verified by conducting interviews and/or surveys with standard-setting participants – the purpose being to ensure that the samples provided for panelists represent a broad spectrum of ability and are representative enough of the communicative demands of the target domain to allow them to decide on the readiness of test candidates. Then, the assumption that “standard-setting panellists are confident in the validity of their judgements” can be supported from evidence of feedback from standard-setting participants. This kind of verification procedure canvassing panellists’ level of confidence in their own decision-making processes is usually sought at the end of a standard-setting workshop in the shape of a final evaluation form. The last assumption is that “standard-setting procedures are suitable and consistently applied”. This assumption can be supported with evidence of a justification for a suitable and appropriate choice of method from the range that are available including a thorough description of the procedures utilized. Feedback can also be gained from standard-setting panelists (e.g., a training evaluation form) to gather evidence on their understanding of the application of the method.
Furthermore, comparisons with other standard-setting studies using the same method can be made to analyze possible similarities and differences in the findings. These kinds of evidence associated with the decisions inference in the interpretive argument for an LSP test (as proposed by Knoch & Macqueen (in preparation)) are sought in the current study and described in further detail in the method chapter that follows.

To sum up, conceptualizations of test validity have undergone significant historical change that points to its complex and diverse nature. Recent argument-based approaches have acknowledged this complexity by spelling out the different kinds of inferences associated with test scores, the various assumptions underlying these inferences and the evidential support that is required to support their validity. This review has outlined the somewhat different formulations and application of the argument-based approach by leading scholars in the field of assessment including the particular application of this approach to LSP testing devised by Knoch and Macqueen (in preparation) that guides the current study.

2.7 Chapter summary

This literature review has offered an account of the published research which has served as a foundation for this thesis. As stated previously, it is commonly agreed in LSP testing when setting standards subject-matter experts should be included in the process because only they possess the knowledge and experience of the requisite communication skills of their own professional sphere. Nevertheless, whether domain experts (as opposed to language experts) are able to establish valid standards should be born in mind as they might consider aspects of performance that are significant to them and yet are not currently captured in the OET Writing sub-test criteria. Uncertainties may persist about the validity of domain experts’ judgements – these questions have rarely been investigated in the literature.

In response, research question one (i.e. What criteria do standard-setting participants use as a basis for their decisions in judging writing responses and to what extent are
these decisions language-based) aims to investigate health professionals’ decisions concerning OET Writing sub-test samples and the performance features that they draw on to make judgements. Research question two (i.e. Is there any variability between judges in what they attend to while setting standards?) attempts to account for any potential variation in panellists’ judgements. Research question three (i.e. How do standard-setting judges view the process and outcome of the standard-setting procedure?) additionally examines any possible variation between participants. Furthermore, by completing final evaluation questionnaires, participants’ overall opinion of the standard-setting procedure is described. A thematic analysis of the qualitative data is carried out to determine relevant aspects that health professionals deem important when judging the satisfactoriness of writing performances (pertinent to research questions one, two and three). In addition, a quantitative analysis is undertaken to inform the foundation of the new standards and associated cut scores. This is in order to answer research question four (i.e. What occupational specific standards (cut scores) do doctors set on the Occupational English Test (OET) Writing sub-test?). As previously mentioned, this study aims to address any potential validity concerns of involving subject-matter experts’ views in the standard-setting procedure by using a distinctive argument-based framework to LSP testing offered by Knoch and Macqueen (in preparation). In addition, an under-used qualitative component employed in this study further investigates any possible validity issues when domain expert participants make performance level judgements. The next section outlines the methods used in this study.
Chapter 3: Method

3.1 Introduction

This chapter contains five sections which describe: 1) the methodological framework that guides the study; 2) a comparison of two potential standard-setting methods for the main study; 3) pilot standard-setting workshops based on these two methods; 4) a review and trial of the think-aloud procedure designed to elicit qualitative insights from participants; and 5) an account of the methods adopted for the main study. The pilot standard-setting workshops and think-aloud trial were undertaken before the main study and informed the final decision of the procedures that were ultimately used. The description of the pilot standard-setting workshops illustrates some of the issues that arose with the two possible methods. It also gives a justification for the final choice of method and procedural refinements that were made. The think-aloud trial provides a rationale for the use of this approach in the main study and an account of some of the practical issues involved. In the main study section, the participants, instruments, research design and procedures for participant recruitment, workshops, verbal protocols (think-alouds) and data analysis are described.

3.2 Methodological approach

This study’s methodology employed a combination of quantitative and qualitative approaches, also known as mixed methods (Brady, Collier, & Seawright, 2010; Collier, Brady, & Seawright, 2010). Creswell, Plano Clark, Gutmann and Hanson (2003) and Creswell and Plano Clark (2007) argue that the use of a mixed-method approach is beneficial when one method does not offer all the evidence required as it includes a number of perspectives. Furthermore, Denzin and Lincoln (1998) note: “The combination of multiple methods, empirical materials, perspectives and observers in a single study is best understood, then, as a strategy that adds rigor, breadth, and depth to any investigation” (p.4). Brewer and Hunter (1989) also state that one of the purposes of a mixed methods approach is to gain further validity in research because
using a combination of methods avoids deficiencies in one practice only. According to Henn, Weinstein and Foard (2009), using a combination of methods or ‘triangulation’ may offer a more valid and ‘holistic’ analysis than if one method alone was employed. Greene, Caracelli and Graham (1989) claim that triangulation in a study purposefully uses more than one method of data gathering and analysis in order to obtain possible corroboration and convergence between the results of different methods. Hence, this study has adopted a mixed-method approach.

A further aim of a mixed-method approach is to achieve ‘complementarity’ (Greene et al., 1989) whereby different methods may be utilised to investigate different aspects of a phenomenon. According to Riazi and Candlin (2014) complementarity is best realized by “carrying out each method interactively/interdependently and concurrently, to cast as much light as possible on the complexity of the research at issue” (p. 144). In this model neither the quantitative nor qualitative data is considered to be the driver of the research design – both are integrated and analysed alongside each other to answer a particular research question. The present study thus takes an ‘interactive’ or ‘equal status’ design (Cresswell, 2009) to its mixed-method approach and the collection of qualitative and quantitative data.

Three sources of data were used in this study: 1) standard-setting panellists’ performance judgements of the quality of writing samples; 2) recordings of workshop discussions; and 3) recordings of think-aloud protocols collected outside of the workshops. The standard-setting method uses quantitative measures for the setting of ‘cut scores’ between performance levels of participants’ responses. Standard-setting workshops were convened with subject-matter experts (medical professionals) for this purpose. Furthermore, to gain additional insight into participants’ decision-making while forming judgements, verbal reports in the form of concurrent think-aloud protocols (TAPs) (Ericsson & Simon, 1993; Green, 1998) were utilized with a sample of participants. A thematic analysis of panellists’ workshop discussions and TAPs was undertaken to discover the foundation for their decisions about appropriate divisions between performance levels. This kind of qualitative study
is an under-researched area in standard-setting studies (as seen in section 2.5.6 of the literature review).

3.3 Choosing a method for the main study

As noted in the literature review, there are many considerations that need to be taken into account when deciding on a method for a standard-setting study (Cizek, 1996; Kaftandjieva, 2004; Reckase, 2000; Zieky et al., 2008). Several standard-setting methods were discussed (see section 2.5.4 of the literature review) that are examinee-centred, focus on test takers’ actual performance and ask panellists to make holistic judgements about assessment performances (Cizek & Bunch, 2007). Other methods also considered whether a test candidate is in a ‘borderline’ or ‘just qualified’ range (i.e. as being just able to meet a particular standard or not) (Tannenbaum & Katz, 2013). In addition, as noted, a number of standard-setting methods are apparent in the literature for setting standards on writing performance tasks (see Cizek, 2001, 2012; Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006; Kaftandjieva, 2004; Zieky et al., 2008).

The choice of method to be used in the main study was narrowed down from a number of previously mentioned alternatives (see section 2.5.4 of the literature review). From the initial consideration of several methods, two were selected to be further compared in two pilot studies conducted at the University of Melbourne: The Analytic Judgement Method (AJM) (Plake & Hambleton, 2001; Zieky et al., 2008) and a modified Performance Profile method (PP) (Tannenbaum & Wylie, 2013; Zieky et al., 2008). These two methods were chosen as they fulfilled a number of key selection criteria (as noted by Cizek, 1996, 2001, 2012; Kaftandjieva, 2004; Reckase, 2000; Zieky et al., 2008). These included: the appropriateness of the method for the particular context; sample size of writing responses; degree of expertise and number of participants; and available resources in terms of time, staff, equipment and funding.

The pilot study took the form of two workshops (see section 3.4). An initial ‘in-house’ workshop with language experts from the University of Melbourne’s School of
Languages and Linguistics (SOLL) was conducted to consider the two possible methods. A second workshop with health professional educators was also undertaken to inform the choice of the final method to be used in the main standard-setting workshops. Outlined below is a summary of the key features of each method as they apply to the focus of this study (i.e. English writing performances). Following this is a description of the two pilot studies, whose purpose was to choose between the two methods. Also, how the two pilot studies were conducted are outlined and the results/insights that emerged from them are offered. Finally, the decisions, implications and amendments that arose from these pilot studies are given for further refinement of the final selected method.

### 3.3.1 Analytic Judgement method (AJM)

As stated in section 2.5.4 of the literature review, in the Analytic Judgement method (AJM), panellists consider a range of test candidate writing performances and form value judgements about these individual writing samples independently (i.e. not with any pre-ordered arrangement of scripts). The AJM requires standard-setting participants to judge and place each sample (depending on the number of cut scores that need to be set) as ‘UNSATISFACTORY’, ‘NOT YET COMPETENT’, ‘COMPETENT’, ‘STRONG’ or in a ‘between’ category. In a training session, panelists are asked to collectively discuss and define each of the main performance levels for which judgements are made (e.g., participants are asked to determine the features that would make a test candidate’s writing sample competent or not yet competent). In a practice round, each participant then receives a small number of writing responses. Panellists study the samples, work independently of each other and assign each response to a performance level using a designated form – see Appendix C (i.e. STRONG, COMPETENT, NOT YET COMPETENT and UNSATISFACTORY or a ‘between’ category – ‘between COMPETENT and NOT YET COMPETENT’ or ‘between COMPETENT and STRONG’). Panellists state their results, discuss any differences and specify why they placed specific responses in particular performance level categories. If there are any areas of substantial disagreement they discuss why. Participants are free to amend their original judgements in light of subsequent discussion on their own
individual form, but do not have to do so. Panellists are then given a larger selection of new responses, again not in any specific arrangement (i.e. randomly). These new scripts have scores that span the full range of those actually obtained by test takers. Once more, participants assign a category independently using a designated form and further allocate the responses to an assigned performance level or a ‘between’ category.

To calculate cut scores, the writing samples’ scores from the ‘between’ performance levels (as categorized by HPs), are pooled and averaged. For example, all the scores for the writing responses that have been placed in a ‘between’ category (e.g., ‘between COMPETENT and NOT YET COMPETENT’) are averaged to produce a final cut score. Participants are not informed of the actual test takers’ scores. Cut score calculations for each performance level are finalized after the participants have completed all their judgements.

### 3.3.2 Modified performance profile method (PP)

For the modified Performance Profile method (PP) (as noted in section 2.5.4 of the literature review), participants in a training meeting, are requested to confer and define each of the main performance levels for which judgements are decided (e.g. participants are questioned about what makes a test candidate’s writing performance ‘minimally competent’ or ‘borderline’). Panellists then receive a pre-ordered booklet of writing responses arranged from weakest to strongest. Test takers’ actual scores are not shown to participants. For the training/practice round, panellists receive a small sample set of test candidate writing responses. Participants study the responses and independently decide on a response as being minimally competent or just qualified for the performance level ‘COMPETENT’. Panellists use a designated form (see Appendix F) and decide if each writing response in the set is ‘Yes or No’ minimally competent. Participants state their results for each writing sample and specify why they placed responses as minimally acceptable or not. If there are any areas of substantial disagreement (e.g., outliers), panellists discuss why and justify their placement.
Participants are then given a larger selection of new responses that are arranged in an ordered booklet, based on scores previously assigned to writing samples from weakest to strongest. Again, the actual score is not given to judges. Once more panellists study the responses and independently decide on a response as being acceptable or not for the performance levels ‘NOT YET COMPETENT’, ‘COMPETENT’ and ‘STRONG’. Participants use a designated form (see Appendix F) and work through the ordered responses and decide if each response is ‘Yes or No’ satisfactory for that particular performance category and where they would ‘draw the line’ between performance levels.

Zieky et al. (2008) note that many facilitators ask participants to begin with the COMPETENT level because the most important distinction for high-stakes decision making is usually between NOT YET COMPETENT and COMPETENT. After participants have completed discussion of each round of performance level under consideration (i.e. NOT YET COMPETENT, COMPETENT, STRONG) cut scores can be calculated. Zieky et al. (2008) advise that facilitators could use the mean, median or trimmed mean of the pooled panellists’ judgements. To achieve final cut scores, the writing samples’ actual scores are used as a basis for the cut score calculation. As noted, participants do not know the actual test takers’ scores and the cut score calculation is carried out after participants have completed all their judgements.

3.4 Pilot standard setting workshops

Two pilot standard-setting workshops were conducted at The University of Melbourne to consider two potential standard-setting methods (Analytic Judgement method (AJM) and modified Performance Profile method (PP)) to evaluate the benefits and drawbacks and any potential issues with each method. As noted in the section 2.5.4 of the literature review on the evaluation of methods, each of the two methodologies under consideration have their own practical and theoretical advantages and disadvantages. The key difference between the AJM and PP method is the arrangement of writing samples. In the AJM the scripts are not presented to
participants in a pre-ordered manner, whereas in the PP method they are in the form of an ordered booklet. This means in the AJM panellists are required to make performance level judgements independently of other writing responses in the set they are allocated. In the PP method, however, participants are aware from the prior ordering that one sample is better than the other. Another significant difference is that in the AJM the judgement and allocation of responses is scalar (on a 7-point scale from UNSATISFACTORY to STRONG) whereas in the PP method the approach is dichotomous (a Yes/No decision for each performance level under consideration). These methodological concerns were investigated in the two pilot workshops and are outlined in the following sections.

3.4.1 Participants

Eight participants from the School of Languages and Linguistics (SOLL) at the University of Melbourne attended an initial ‘in-house’ presentation to compare the relative merits of each method in relation to the aims and circumstances of the current study. In addition, a second pilot workshop was conducted with two health professional educators (P1 and P2) with a background in medical/nursing education from two Melbourne universities. They were also members of the overall ARC-funded research project team. The purpose of this second study was to include people who were actually in a position to make judgements about cut scores as it was not true for the previous group who lacked the relevant professional experience.

3.4.2 Materials and procedures

In the initial in-house session, the participants were introduced the ARC project, an outline of the OET and its purpose and a standard-setting overview. Participants were then presented with a comparison of the two methods: Analytic Judgement method (AJM) and modified Performance Profile method (PP). The workshop members were then asked to judge a small selection of OET test candidate writing responses using the two methods (two samples for the AJM and three for the PP). Subsequently, the two method’s procedures and calculation of cut scores were evaluated and
considered. Participants discussed the following questions: 1) which method of presenting the writing samples did you think was better; 2) which approach to eliciting judgements did you find easier to understand/use; 3) which cut score calculation was easier; and 4) which response form/s and alternatives (e.g., the AJM’s 7-point or 12-point scale – see Appendix D and E) did you think would be easier for participants to understand/complete?

In the second pilot workshop, the two HP participants were given a similar introduction to that given in the previous initial presentation. This included an overview of the ARC project, the OET and standard setting. HP participants were asked to once more judge a range of OET test candidate writing samples using the two methods (three scripts for the AJM and four for the PP). The workshop attendees again considered the two methods’ procedures, forms and cut score calculations and their comparative advantages and disadvantages. The HP participants discussed similar evaluative questions as mentioned above in the first pilot. The pros and cons of each of the considered methods are further outlined in the following section.

3.4.3 Results of the pilot workshops

This section concurrently examines the two potential methods, the Analytic Judgement Method (AJM) and a modified Performance Profile Method (PP), that were under consideration for the main study. The issues that were raised from the two methods in each pilot workshop are discussed in parallel. Drawing on the insights gained from this discussion, a final choice of method was made for the study’s principal standard-setting workshops and the justification for this choice is outlined. The following discussion is summarized under headings representing the key themes that emerged from the feedback.

3.4.3.1 Ordering of writing responses

As mentioned, the AJM requires standard-setting participants to form value judgements about individual writing responses independently (i.e. not with any pre-
ordered arrangement of scripts) and place them as UNSATISFACTORY, NOT YET COMPETENT, COMPETENT, STRONG or in a ‘between’ category. During the language professional presentation, it was remarked upon that the AJM method may lend itself better to the standard setting study’s goal of setting revised standards, as it is based on HP’s independent judgements without the possibility of being influenced by the scores assigned to scripts by language raters. HPs use their ‘own’ value judgements about minimal levels of language proficiency under entry-level supervision based purely on their ‘own’ knowledge and experience in the workplace.

In comparison, a possible concern with the PP method is that participants are already explicitly aware that the writing samples have been pre-ordered from weakest to strongest. This may not be as sound methodologically because panellists already have implied knowledge of a writing sample’s ‘actual score’ and therefore in making a judgement about whether a script is minimally competent or not, they may spend less time assessing the writing response if it was not already pre-arranged and they had to evaluate it on its own merits (as in the AJM). Language professionals argued HP panellists might take a more ‘cursory’ approach to judging individual scripts as they already have a view from the response ordering that one is ‘better’ than another.

HP workshop participants (P1 and P2) also favoured the AJM over the PP method for this reason. P1 stated that in relation to the PP method: “I found it hard to determine where that cut off was. Yeah, I prefer the other way (AJM)” and also P2 “I preferred the first method (AJM)”. P1 elaborated that:

Because I think you can just look at something and think well…it just enables you to have that gradation approach. With the second method (PP) it’s just yes and no and it’s a bit more severe especially when you look at something and you think this is meant to be a really bad one, but actually it’s not too bad. And you are forced to make a yes or no decision and I think that’s really hard rather than being able to be given a choice over a range of samples.

As shown, a concern with the PP, noted in the language expert presentation, could be a panelist might disagree that one writing sample in the lowest to highest arrangement is better than one lower down in the ordering. If a participant had an
issue with the ordering of the booklet it might be a barrier to making judgements about writing samples. A HP may feel distracted if his/her opinion differs about what the ordering from weakest to strongest should be or it could lead to questions about the validity of HP’s judgements if their intrinsic ordering is different to English-language professionals. P2 from the HP workshop confirmed this issue by stating:

And also, the fact that it’s not only a yes or no it’s a yes or no on a previously looked at sample and somebody else has made the decision so you’re sort of going...sometimes someone might get ‘bolshie’ and say I don't agree with the ordering of samples. In other words, I might disagree with the ordering and think that the one I just read is better than the previous one or maybe worse.

Consequently, there are different levels of engagement with each script depending on the method. With the AJM, participants have to engage individually with each script because they are independent of each other; they have no idea what they were pre-rated as and the next one they read could be much stronger or weaker compared to the previous one. Therefore, panel members arguably need to engage more with each writing sample in the AJM compared to the PP method.

The issue of whether HP panellists’ judgements are to be guided by linguistically predetermined ratings is an important one. A concern is whether participants should be more independent in their thought processes and forming of judgements which are then later linked back to the OET. The PP method’s limitation is that the judgement of samples is already framed and pre-ordered in terms of the OET rating scales, and this may interfere with the workshop’s purpose of allowing HPs to give voice to their indigenous assessment criteria (i.e. those by which they would normally judge the writing samples independent of any other considerations such as what the language trained raters might value about the writing). Thus, conceptually the AJM may be more suited to the standard-setting workshop’s purpose.
3.4.3.2 Qualitative data elicitation

A related issue is the type of method that lends itself best to rich commentary for qualitative thematic analysis. In regard to qualitative data, P1 from the HP workshop stated:

The AJM is better for eliciting comments and feedback from participants because then you can qualify why it is a STRONG or COMPETENT sample rather than just a yes or no so you’re more likely to get richer data because you have to justify it with a lot more evidence rather than if it was just a yes or no.

As noted, AJM participants have to evaluate each script independently of other samples and make completely new and separate judgements rather than the PP in which more ‘superficial’ judgements might be generated. Therefore, the AJM may be better suited to generating insights into the reasons underlying participants’ judgements. The focus on qualitative data associated with standard-setting judgements is a key component of this study and an under-researched area of investigation in standard setting.

3.4.3.3 Conditions for data elicitation

3.4.3.3.1 Format for judging samples

All workshop participants evaluated the format for judging samples. HP participants preferred the scalar AJM to the dichotomous PP method. P2 from the health professional workshop illustrated the reason for this as:

Because you’re looking at it for all the different points rather than ‘is it competent or not?’ I quite like the idea of trying to decide whether they’re good, bad or indifferent, not just competent.

Two versions of the AJM’s categorization of performance levels were presented to both the language professional and HP professional trial workshops. As noted in the literature review (see section 2.5.4), the AJM in its original format uses a 12-point categorization of performance levels. The performance levels, for example, STRONG, COMPETENT, NOT YET COMPETENT and UNSATISFACTORY are further divided into
high, medium and low. This is how the method was originally conceived by Plake and Hambleton (2001). The highest-level category is therefore ‘HIGH STRONG’ and the lowest is ‘LOW UNSATISFACTORY’. The setting of cut scores is achieved by focusing on the borderline categories. For instance, for the adjoining performance levels of ‘HIGH NOT YET COMPETENT’ and ‘LOW COMPETENT’, all the writing samples that participants put into these two categories are used to determine a single cut score between NOT YET COMPETENT and COMPETENT. This is done by calculating the mean of the scores that were already provided using a predetermined rating method. As noted in the literature review (see section 2.5.4), Plake and Hambleton (2001) put forward an adaptation of the AJM in which the borderline groupings are combined to form a set of seven categories. Cizek and Bunch (2007) argue that seven rather than 12 categories was considered more manageable and easier for participants to conceptualize performance level distinctions. As noted, a seven category AJM version was used successfully by Pill and McNamara (2016).

In regard to alternate AJM variations/forms (see Appendix D, E) and the number of performance level categories, the two pilot sessions considered both the seven and 12 category versions of the AJM. The seven-category adaptation was preferred by HP participants. P2 stated “it’s probably easier and for group standard setting to do seven categories rather than 12”. The 12-category version was considered problematic by P1 from the HP trial workshop “...because it’s hard to decide between something that’s high, medium and low from COMPETENT and NOT YET COMPETENT and is a bit subjective”. Therefore, the seven-category version of the AJM was thought to be preferable overall.

A further concern with the AJM raised in the language professional trial is there may be participants who might be hesitant to place samples in ‘boundary’ or ‘between’ classifications as they do not like to ‘sit on the fence’. Therefore, it needs to be stressed to panel members that the ‘in-between’ categories are ‘real’ categories and that they are to be used as freely as the others. Also, a consequent concern with the AJM is that there may not be enough samples classified in the ‘boundary’ or ‘between’ categories of performance levels and hence it may make it problematic to calculate
cut scores. This type of issue would not occur with the PP method as participants need to only decide for each sample whether it is ‘Yes or No’ to be included or not in a particular performance level under consideration.

### 3.4.3.3.2 Number of participants

Concerning the number of participants for the study, Zieky et al. (2008) (as noted in the literature review) suggest that overall 12-18 participants be used (although this number is arbitrary). P2 claimed it would be “better to do it in groups as you’d probably get better data”. Also, P1 suggested “some may feel intimidated or might just go along with the group consensus like ‘oh yeah what she said and that’s ok’ so they’ll just accept it”. Therefore, small groups might be preferable. The issue of the number of participants in a study also overlaps with the independent judgement question mentioned previously. A participant in a smaller group may be more inclined to make more independent judgements than if they were in larger group and felt pressure from the other group members. However, group consensus is not an aim of either the AJM or PP method.

### 3.4.3.3.3 Number of prompts/scripts

In the language professional trial workshop, it was noted that the AJM method might be more time consuming for participants compared to the PP and hence fewer writing responses could be included and judged. A solution might be for some writing samples to be taken home and judged post-workshop. In contrast, in the PP, the method’s procedure might allow more scripts to be seen in a given time frame by panellists. This is because the decision-making process of judgement is arguably easier whereby participants hone in on ‘Yes or No’ classifications of minimal competency.

When discussing how many samples would be acceptable for participants in an AJM standard-setting session, HP participants recommended that for the training workshops there should be three or four samples with discussion of each and actual samples of around 15. P2 stated in relation to a recent standard-setting workshop:
“We did 20 exam papers the other day in 1.5 hours and by the end we were bored and a bit grumpy so I reckon 15 might be the limit”. Hence, the number of possible scripts needs to be considered as manageable for participants.

3.4.3.3.4 Time management

As mentioned, P2 had participated as a standard-setting panellist for another unrelated study and offered some overall advice in running the workshops more efficiently in terms of time. In regard to the procedure of conducting a standard-setting workshop P2 stated:

The other way of doing it is, as we do them, do them and send them [performance level judgement forms] in. So that way you save time and really get them to do them by themselves. You get a lot more discussion then. So, you get them in, train them, send them away and get them back in a week later to just discuss. And then the discussion is really quite interesting. And it actually would be shorter because you could do it in an hour.

However, P1 argued: “You have to bring them in twice which is another issue because you have to train them and send them off and then bring them back again”. Yet, P2 contended that “participants might rate scripts more carefully if done individually at home and more superficially in a workshop as they’d be against the clock so to speak”.

3.4.3.4 Issues common to both methods

Some shared concerns were raised that might be problematic for either of the two methods during the ‘official’ standard-setting workshops. Firstly, both language and HP trial workshop participants noted that HP panellists in the main study’s workshops may argue that the OET Writing test task does not accurately reflect the writing demands of their workplace or that it is not a task that is regularly performed by entry-level staff. While task authenticity and validity are not a central focus of this study, participants’ reaction to the task and the extent to which this may influence their judgements regarding standards is discussed later in this thesis.
Another concern that was raised by all trial participants is that some health professionals find it challenging to make the distinction between linguistic and professional competency while judging performance samples. This was also noted by Pill (2013) and Pill and McNamara (2016) when panellists were judging speaking samples. Therefore, participants in the main study need to be reminded regularly both verbally and in written form in the training materials that the standard-setting procedure is concerned with language proficiency rather than clinical competence. As noted, the Australian federal government requires that the assessment of language be separated from the assessment of clinical competence.

### 3.4.3.5 Final choice of method

Two potential methods, the Analytic Judgement method (AJM) and a modified Performance Profile method (PP) method were considered for use in the main study. The results, themes and issues with each method from two trial workshops at the University of Melbourne were highlighted. From this, a decision on the choice of a final method was reached. The AJM was chosen as the final method to be used with health professional workshops in the main study. The decisions for this are summarized as: 1) conceptually the AJM is better suited to the standard-setting workshops’ aim. The PP’s pre-ordering and framing of writing samples in terms of the OET rating scales might interfere with the purpose of the study (i.e. to elicit health professionals’ ‘independent’ judgements of performance standards based on their professional knowledge and experience); 2) participants preferred the scalar 7-point version of the AJM rather than the dichotomous ‘Yes or No’ judgement approach of the PP method; 3) greater levels of engagement might be demonstrated with each writing response in the AJM compared to the PP method; and 4) the AJM may be better suited to eliciting qualitative data (an underexamined aspect of standard-setting research).
3.4.3.6 Procedural refinements to the chosen method

A number of practical enhancements to the running of the standard-setting procedure using the AJM emerged from the pilot sessions. Firstly, as the time needed to judge writing samples may be greater using the AJM in comparison with the PP method it was decided that a pre-task should be sent before the training workshop and also that writing samples could be taken home and judged post-workshop. This would increase the amount of data on which standards would be calculated and hence the representativeness and robustness of these standards. Due to time, budget and logistical restraints it was not possible to ask participants to return and discuss their performance level judgements as suggested by P2 in section 3.4.3.3.4. However, pre- and post-training workshops would help alleviate these concerns. Regarding how many samples could be discussed in the main study’s training workshops of one hour, HP trial participants recommended three or four scripts, but with a pre-task this could be increased to around five. Also, HP trial participants suggested that if HP workshop panel members were to take home and judge scripts, the total number of samples could be increased. Therefore, a broader range of responses could be seen and rated overall. Concerning the alternative AJM forms and the number of performance level categories to use for decision-making purposes, HP trial participants recommended seven categories rather than 12. Finally, to avoid the risk of HP panel members misunderstanding the intended focus for their standard-setting judgements, it was deemed important that the OET’s purpose of assessing communication skills rather than clinical competence should be highlighted as part of the briefing to participants.

3.5 Qualitative data collection methods

Many standard-setting researchers have stressed the necessity to try to comprehend the cognitive processes motivating participants’ behaviour (see e.g., Buckendahl, 2005; Giraud & Impara, 2005; McGinty, 2005; Plake & Impara, 1997). Numerous data collection methods have been employed to analyse standard-setting participants’ thought processes while they make judgements. These methods are usually in verbal or written form such as interviews, focus group discussions and ‘think aloud’ protocols.
(Cizek & Bunch, 2007; Zieky et al., 2008). This section discusses the qualitative data collection methods utilised in this study.

3.5.1 Group discussions

Deciding on performance level judgements involves standard-setting procedures that employ knowledgeable subject-matter experts to make independent judgements, note differences and specify the degree to which agreement was reached (Kenyon & Römhild, 2014). Hence, it is usual for standard-setting panels to involve multiple participants and some form of discussion between panellists concerning the performance standards under review. Some panels though may be convened via other arrangements such as online (e.g., see Katz, Tannenbaum & Kannan, 2009; Katz & Tannenbaum, 2014). All standard-setting panellists in this study participated in a standard-setting training workshop. The workshop discussions were conducted using “open-ended questions and informal probing to facilitate a discussion of issues in a semi-structured or unstructured manner” as this is important “when seeking to understand people’s motives and interpretations” (Devine, 2002, p. 198).

A number of advantages are evident in group discussions such as a range of views can be canvassed quickly, panellists may provide checks and balances on each other to limit extreme or outlier views and the extent to which there is a relative consistency of views or great diversity can be evaluated relatively easily. However, group discussions may also have some limitations. For example, a participant with an alternate viewpoint or less vocal personality might feel less inclined to speak out and face a negative reaction from other group members. In addition, the discussion may be hindered or shortened so that all participants are able to be heard in the allocated time-frame (Patton, 2002). Therefore, other complementary methods in addition to group discussions, such as those stated below, may also be useful to understand standard-setting informants’ thought processes.
3.5.2 Structured interviews

Some of the participants were asked to participate in a short, structured interview. This interview was part of a final evaluation of the think-aloud protocol (TAP) stage of the study – described in detail below in section 3.6.4.4.2. All participants were asked questions that intended to elicit evidence to answer research question three (i.e. How do standard-setting judges view the process and outcome of the standard-setting procedure?). The reason for choosing a structured interview approach (in comparison to a semi or unstructured method) is due to: 1) the questions were designed to confirm participants understanding of the standard-setting and think-aloud (TAP) tasks and procedures and a lengthy discussion was not required; and 2) the limited allocated time frame available to health professional panelists did not allow for an unstructured or extended interview.

3.5.3 Think-aloud protocols (TAPs)

In this section, firstly background to think-aloud protocols (TAPs) is stated; secondly, advantages and potential issues with using TAPs are given; and lastly a TAP trial is described.

3.5.3.1 Background

In preparation for the main study, it is appropriate to review the use of verbal reports such as think-aloud protocols (TAPs) and consider their potential role in the qualitative data collection process. A TAP is a technique that has its origins in cognitive psychological research (also used in a variety of social sciences) that were initially used to examine mental processes such as problem solving (Ericsson & Simon, 1984, 1987, 1993). An informant is asked to ‘say out loud’ their thoughts as they carry out a specific task and verbalise anything that comes into their mind such as what they think or notice. Therefore, a TAP may allow a researcher to gain insight into a participant’s explicit cognitive processes while carrying out a task. These verbalizations are
The use of TAPs in educational research and applied linguistics is becoming more widespread (see e.g., Cohen, 1987, 1988, 1996, 2000). They have been utilized extensively to examine the process of essay rating in first language (L1) contexts (see e.g., Huot, 1993; Wolfe, Kao & Ranney, 1998) and second language environments (see e.g., Cumming, Kantor & Powers, 2002; Lumley, 2005). However, their widespread utilization in standard-setting studies is not yet evident. An exception is Dawber, Lewis and Rogers (2002) who used a TAP and a questionnaire to investigate the thought processes of participants and their understanding of the ‘Bookmark’ standard-setting method. Their study considered three standard-setting committees and demonstrated that their understanding of the standard-setting procedure changed between rounds of judgement. Despite a TAP being employed in this particular study, the reason why they aren’t used more extensively in standard-setting contexts is unclear. A possible explanation may be as Skorupski and Hambleton (2005) state in their qualitative-focussed standard-setting study “their preference was to have panellists ‘think aloud’ to express their thoughts about the process” but that “such an activity would likely have been disruptive to the process of operationally setting standards” (p.234). Participants were instead instructed at various stages (pre, during, post) of the standard-setting workshop to write down their responses to key questions concerning their feelings and thoughts about the procedure. It is arguable whether asking participants to stop and write down their thoughts is more or less distracting than utilizing a TAP. With a large group of panellists present, as is often the case in standard-setting sessions, this type of disruption to the procedure is likely to be true. The advantages and disadvantages of using TAPs are now be discussed.

3.5.3.2 Benefits of using TAPs

There are a number of benefits of employing TAPs in qualitative data collection. The first advantage, as opposed to other qualitative methods such as interviews or questionnaires, is the immediacy of the approach. The data obtained by using a TAP
has an added proximity to the focus of enquiry (Ericsson & Simon, 1984, 1987, 1993). Hughes and Parkes (2003) state that concurrent think-aloud protocols may offer “a real-time insight into the knowledge that a subject uses and the mental processes applied while performing a process of interest” (p. 127). Brown (1987) also argues that TAP methods may be able to offer more authentic accounts of participants’ thought processes and behaviour instead of sweeping statements taken from numerous incidents or occurrences, as in interviews or questionnaires. In addition, Vann, Lorenz and Meyer (1991) contend that interviews and questionnaires are capable of only supplying data about participants’ consciously understood thoughts, which may or may not express how they ‘actually’ think. Therefore, TAPs may be able to provide further understanding of participants’ thought processes than post standard-setting workshop questionnaires and interviews alone. The second advantage of using TAPs is the directness of the method. By examining thought processes as they happen, TAP data may offer insights that are not generated by way of other types (i.e. questionnaires or interviews). Cohen (2000) claims that verbal reports present information on cognition “that otherwise would have to be investigated only indirectly” (p. 129). Furthermore, DeRemer (1998) argues that only by analysing participants’ verbalizations during a rating session can comprehension of how judgements are formed begin to be made.

### 3.5.3.3 Issues with TAPs

There has been some criticism in regard to the validity and consistency of data gathered using TAP methods and these issues are discussed in this section. One of the early concerns is TAP data’s claim of ‘direct’ evidence of cognitive processes. Initial criticisms in cognitive psychology argued that direct confirmation of cognitive processes is unachievable (Cooper & Holtzman, 1983) and there is no apparent connection between a participant’s cognitive processes and their verbal report data (Nisbett & Wilson, 1977). In response, Cohen (1987) observed that verbal reports may merely ‘inform’ on the conscious strategies participants use. Furthermore, Bracewell and Breuleux (1994) conceive that TAPs are only a sign of cognitive processes as a kind
of “trace data” (p. 85). Ericsson and Simon (1984, 1993) conclude then that TAP data merely ‘indicates’ rather than directly exemplifies thought processes.

Other data validity issues also remain concerning research design. The first area of concern is the time interval between the chosen task and the reporting on that task. Many researchers in the literature agree that ‘concurrent’ compared to ‘retrospective’ reports offer data with greater levels of completeness and accuracy (Cohen, 2000; Ericsson & Simon, 1984, 1987, 1993; Mackey & Gass, 2005). Therefore, concurrent TAPs are advised in contrast to retrospective ones (Ericsson & Simon, 1993; Cohen, 2000). The second research design issue concerns the nature of instructions given to TAP informants such as the degree and type of prompting which may have consequences for data validity and reactivity concerns. Cohen (2000) differentiates between ‘self-revelational’ and ‘self-observational’ verbal reports which refers to instructions where participants simply verbalise their thoughts and the later where informants are required to include further explanation or analysis. In this regard, participants should only speak out loud about aspects that they would only usually take note of or this could add an extra cognitive burden and hence change the usual performance of the task. Ericsson and Simon (1984, 1993) state that another significant consequence for TAP data quality is how specific the instructions given to participants should be. They argue that more valuable data will be produced if informants are given instructions that are less specific and more general and for participants to use a verbalised interior monologue of whatever comes into their mind. Additionally, if a participant stops talking during a think-aloud report Ericsson and Simon (1987) recommend that the researcher’s prompts be impartial and neutral and should not have an impact on the informant’s focus.

Another issue to contend with is making sure that TAP data is ‘complete’. In an empirical study on the use of TAPs in essay rating, Barkaoui (2011) investigated ‘veridicality’ issues and if TAPs are able to give a true account of a participants’ thought processes and thinking overall. The study’s results showed that TAPs “are necessarily incomplete and likely to alter the rating process” (p. 70) and that the findings need to take into account any inherent limitations. Hence, it may be necessary to complement
TAP data with other sources. It is suggested then that the use of numerous kinds of information will afford greater faith in the data collection procedure (Jourdenais, 2001). Ericsson and Simon (1987) also argue that extra sources of data can allow “convergent validation” of TAPs (p. 51). For instance, Milanovic, Saville and Shuhong (1996) used concurrent verbal reports with group interview data and retrospective written reports in order that the methods could counteract the shortages of each technique.

As seen in this section on think-aloud protocols, despite some concerns with some potential issues, a number of research design measures may be put into practice to attain valid and reliable qualitative TAP data. To summarise, firstly, TAPs should be conducted concurrently to attend to data about thought processes from the present situation instead of from memory. Secondly, instructions should be general to facilitate participants’ ‘natural’ responses and should be devised to promote non-metacognitive data (i.e. without requiring further analysis or explanation). Thirdly, complementary data from other sources should be used to give convergent confirmation of the verbal report.

3.5.3.4 TAP trial

A trial of the TAP (Ericsson & Simon, 1993; Green, 1998) procedure was carried out with the same health professional educators P1 and P2 from the comparison of the standard-setting methods trial (see section 3.4). The participants offered their thoughts in a concurrent TAP while reading and judging ten OET test candidate writing responses. The aim of the trial was to test the procedures of the think-aloud method for later use in the main study. This included considerations such as the maximum number of responses that could be judged comfortably by participants in the allocated 45–60-minute time frame. Both participants were also asked for their reactions immediately following their think-aloud session using a structured interview approach. Feedback was sought on the effectiveness of the TAP instructions that would be given to participants both initially before they started the procedure and the level of instruction and prompting from the researcher that might be required during
the participant’s actual think-aloud report. From the trial, the procedures for the main study were further refined.

### 3.5.3.4.1 Materials and procedures

In the verbal report trial, both participants read and judged the same ten OET test candidate writing responses using a think-aloud protocol (TAP) (Ericsson & Simon, 1993; Green, 1998). The writing samples were chosen to represent a cross-section of performance levels. To compensate for any possible ‘order effect’ the writing responses were arranged and presented randomly to each participant (Sakyi, 2000). An order effect can be observed where a readers’ opinion of the quality of the response might be swayed by what has been previously read (Milanovic et al., 1996) so randomizing the order for different raters mitigates this effect. Both participants were sent training materials before attending the think-aloud session. These materials included information about what a think-aloud involves and some general instructions about the procedure (see Appendix G). These were very similar to the written and verbal instructions given to participants immediately prior to beginning their verbal report (as seen in more detail below). The training materials also included a link to a video (https://www.youtube.com/watch?v=nJ2udLjdsx4) that demonstrated an example concurrent TAP in which a doctor commented on the usability of a medical website.

The participants were not provided with any set criteria for making performance level judgements. They were however reminded to use their own criteria that had been voiced in the previous training workshop. These came from participants’ discussion on levels of ‘minimal’ competency for written communication in a healthcare setting and what would make a test candidate’s writing response STRONG, COMPETENT, NOT YET COMPETENT and UNSATISFACTORY. The raters were asked to read the ten OET Writing task responses and allocate each to a performance level category. While doing this they were instructed to speak out loud their thoughts into a digital recorder and give a verbal report of their thoughts while judging and allocating a response to a
particular performance level. The rating sessions were conducted individually with each participant and took approximately 45 – 60 minutes.

Participants were first instructed to familiarise themselves with the case notes and task (Task 1, Brian Edwards – see Appendix A). They were reminded of the Analytic Judgement method (AJM) procedure that they had encountered in the previous comparative methodology trial and asked to review the check box form (see Appendix H) and the AJM performance level categories. Following this, each participant was asked to read the think-aloud instruction sheet that gave written guidelines on the TAP procedure (see Appendix H). This instruction sheet utilized the developmental work of Cohen (2000) and Ericsson and Simon (1993). These written instructions were also reiterated to participants verbally. Participants were asked to give an overall rating of the written communication skills in each response concurrently while reading each and to say all their thoughts ‘out loud’.

Further to these general instructions, participants were asked to: 1) include comments on what aspects captured their attention and the factors they took into consideration in deciding on a performance level; 2) give as much detail and elaborate by pointing out examples and anything specific that contributed to decisions; and 3) keep up a steady ‘stream of consciousness’ without stopping to think too much about what they would like to say. Each participant was then asked if they understood the written and verbal instructions, had any questions about the TAP procedure or what was required of them during their verbal report.

Throughout the TAP trial and to reduce the risk of interruption, prompts used by the researcher were ‘non-mediated’ (Green, 1998). If a participant’s ‘stream of consciousness’ during their verbal report breaks, a prompt such as ‘keep talking’ is recommended. Such prompts proved to be largely unnecessary as the two trial participants kept up a constant flow of talk with limited pausing. However, on occasion, ‘mediated’ prompts such as ‘can you be more specific?’ (Green, 1998) were used to clarify or elaborate on particular occurrences where the participants’ use of a referent or their meaning was not clear. Another form of mediation on the part of the
researcher was to ask the participant to give an example to a particular referent such as ‘can you highlight any examples?’ Also, sometimes participants were asked: ‘Is there anything else?’ This was usually done at the conclusion of speaking about each writing response and to finalize their thoughts and overall performance level judgement.

Using mediated prompts does involve a possible risk of altering how the task is usually completed by a participant (Ericsson & Simon, 1993). For example, a participant may for the benefit of the researcher, address an aspect of the response that they might not have if they had not been prompted. This may pose a possible validity issue as far as the elicitation of truly ‘indigenous’ criteria are concerned. Elder and McNamara (2016) found discrepancies between obvious mention of indigenous criteria from physiotherapist health professionals. In a training context, at an authentic workplace situation, HPs were not as explicit in their commentary. However, clinical educators in a research-based workshop environment were much more explicit. The risk of using mediated prompts is therefore at least partly offset by the benefits of such instigation which potentially makes more information available than might otherwise be the case if the participant had not been prompted. Green (1998) points out that the use of mediated prompts and the extent of their risk is not conclusive and that some methods of implementation might be less susceptible to intrusiveness than others. For instance, mediated prompts can be worded so that their possible interference might be mitigated, such as in the examples above (Green, 1998).

### 3.5.3.4.2 Post-session interviews

A post verbal report interview was conducted with each participant immediately following their TAP session and rating of OET Writing sub-test responses. This was done using a structured interview approach. The interviews lasted approximately ten minutes. Participants were asked to speak about the TAP procedures and their feelings about the experience of ‘thinking aloud’ overall. The interview questions concerned whether: 1) the instructions were clear and the participant understood what they were required to do when ‘thinking aloud’; 2) the level of prompting from the
researcher was adequate; 3) the panellist was comfortable with the number of writing responses they judged and spoke about; 4) they had adequate time to speak comprehensively about each writing response; 5) the process of ‘thinking aloud’ altered the way in which they judged the writing responses; and 6) they had any further comments or suggestions to enhance the think-aloud procedure.

### 3.5.3.4.3 Results of the TAP trial

The results of the TAP trial and following interviews had direct consequences for the procedural development of the sessions in the main study. Participants’ comments in relation to the structured questions presented above were analysed according to theme. These themes are presented in the section below.

In regard to the quality of the written and verbal instructions, both participants found no issues and said that they were clear and easy to understand. P1 stated “the instructions were good” and P2 “fine, very clear”. The participants were asked about the level of prompting and whether they considered it to be intrusive or not. In reference to how much or little prompting was thought necessary, P1 said:

You were actually prompting and I thought that was good. It wasn’t just like you were completely quiet. I actually like someone who does talk and I think the level you gave was quite good; it wasn’t too much or too little.

Also, concerning the level of ‘non-mediated’ prompts and how intrusive or not they were, P2 stated that jumping in with a prompt such as ‘can you be more specific’, if a participant is not being 100% clear, “would be useful”. P2 was also asked if such a prompt would interrupt the flow of their verbal report and responded with: “I wouldn’t have thought so”. P2 went on to say that asking a participant to further clarify a point or example might be necessary: “I think if it’s something that you know, we seem to be going off on a track that isn’t connected, then I think that’s probably reasonable to say could you explain a bit more there”.

121
Regarding the order of the written/verbal instructions and the case notes/task given to participants, P2 suggested that the present ordering be switched (i.e. previously each participant was given the case notes/task first and the written/verbal think-aloud/AJM procedure instructions second). P2 advised that it would be “better for participants to read the instructions first, then this [the case notes and task] just so you’ve got the clinical stuff in your mind, even though they can refer to it”. Participants were told that they should not take into account a test candidate’s clinical knowledge while judging writing responses because, as noted, it is a legal and professional registration requirement of the Australian government that clinical knowledge and language proficiency be tested separately.

Regarding the number of scripts that the trial participants thought would be acceptable and comfortable for the main study’s think-aloud phase, both agreed that 12 responses would be satisfactory. This was two more than was originally anticipated as tolerable for participants in the 45 – 60 minutes’ time frame. P1 stated: “I think 12 is good and they’re not too lengthy so they should be able to read it and make some analysis. I think that’s reasonable. I think that’s good”. P2 further corroborated that “12 would be enough in an hour” as the number of responses that would be deemed adequate for participants to comment on. In terms of the trial participant’s thoughts on the TAP procedure overall, neither of them had any issues or problems with the process. P1 said “I thought it was all fine. It was good” and P2 reiterated “the process itself was easy, but yeah I like talking out loud”.

In summary, both trial think-aloud participants believed that the written and verbal instructions were sufficient and easy to understand. The ordering was modified for the main study so that participants received the written/verbal instructions first and the case notes/task second. In regard to the level of prompting, the use of non-mediated prompts was not needed as the participants in this trial didn’t make lengthy pauses. However, in the main study, the suggestion of Ericsson and Simon (1987) that prompts should be given after “15 second to one-minute pauses” (p. 37) (as appropriate) was followed. Concerning mediated prompts, these were deemed necessary at times (and this was confirmed by the trial participants’ comments). In the
main study, their use was determined as appropriate and required. The number of
responses that participants judged in the main study was increased slightly from ten
to 12. This increased number is in line with what the trial participants regarded as
feasible in the 45-60-minute time frame and did not preclude them saying all that they
felt they needed to say about the responses.

3.6 Main study

This section presents the methods utilized in the main study. First, the participants,
ethical clearance and instruments used are introduced. Second, the procedures are
outlined including: participant recruitment, writing response distribution, writing
response allocation to participants and workshop procedures (training and think-
aloud protocol (TAP) session). Next, the data analysis procedures (qualitative and
quantitative) are illustrated. Finally, an argument-based validity framework and
related sources of evidence obtained in this study are described.

3.6.1 Participants

The selection of participants is an important consideration for convening any
standard-setting panel and a number of factors need to be taken into account. Zieky
et al. (2008) assert the most significant aspect is panellists need to be qualified to
make a decision on the required level of skills and knowledge being assessed by the
test to set a cut score. In terms of numbers of participants, Zieky et al. (2008)
recommend that the more contentious a cut score study is likely to be, the more
panellists will be required. However, in general they suggest 12-18 participants should
be used per panel (though they also admit any such number is arbitrary). If this is not
feasible or achievable, they advise the minimum number of participants should be
eight for a study that is used to make consequential decisions (Zieky et al., 2008).
Demographic characteristics such as the composition of men and women should also
be considered for a standard-setting panel.
The 18 participants in this study were drawn from a wide range of health professions. These included general practitioners (GPs), specialists, consultants and medical educators. It was advantageous that the health professionals had a variety of specializations, affiliations and years of experience in the workplace in order to consider the results as broadly as possible. This meant there were medical practitioners who were involved in different workplace activities or clinical specialties and with varying levels of authority and supervision. Zieky et al. (2008) state that many standard-setting studies recommend a criterion of “years of relevant experience” (p. 46). Therefore, participants recruited for this study were required to have had at least a minimum of two-year’s work experience since professional medical registration. The participants had a minimum of four and a maximum of 42 years of experience since registration. The average number of years since registration was 21 (see Appendix I). This indicates the participants had the required levels of experience to make valid and considered standard-setting judgements. It was also deemed important to have panellists who had former exposure in their workplace either as clinicians or educators to the format, style and relevant content features of a referral letter (the type of writing sample that the OET written component utilizes). All participants had experience writing and receiving referral letters in their professional role. Furthermore, it was considered desirable that some panellists had supervisory experience as these participants may have been able to offer guidance and advice to newly registered doctors on the requirements and standards of written communication needed for safe and effective workplace practice. Just over half (11 out of 18) of the panellists had supervisory experience ranging from six months to 30 years of experience. The average length of work experience was 13 years (see Appendix I). This again shows that the study’s participants had the necessary degree of supervision of entry-level doctors to make legitimate standard-setting judgements.

In addition to general supervisory experience, it was also thought to be advantageous to include panel members who had some contact with non-native English speakers in the workplace (suitably with a managerial role). This was true for just under 50% of participants (eight out of 18) (see Appendix J). Furthermore, it was preferable that panel members should have had at least some form of interaction with non-native
English speakers/writers in the workplace. Half of the participants (9 out of 18) had ‘moderate’ to ‘high’ workplace interaction with non-native English-speaking colleagues. The remaining panel members had ‘limited’ interaction (see Appendix K). Furthermore, participants should ideally have seen evidence of non-English-speaking background (NESB) workers’ referral letter writing in their professional context. 50% of all participants said they only had ‘limited’ experience reading referral letters (or similar pieces of communication) written by a NESB doctor per month (0-2 letters read per month). Four out of 18 panellists had read 3-5 letters, three out of 18 had read 6-10 letters and two out of 18 had read 10+ letters per month written by a NESB medical practitioner. Half of all participants had ‘moderate’ to ‘high’ levels of experience reading referral letters written by a NESB doctor. This shows that in general, most panel members had some interactional and referral letter reading experience with medical practitioners from a NESB (see Appendix L).

The current study recruited 18 participants overall which included five men and 13 women. While there was a disproportionate number of females in the sample, it was considered that this would not have a significant influence. Participants in the study were represented by a number of countries of birth, native language backgrounds and countries where the majority of medical training was undertaken. It was thought to be desirable to include participants whose first language was not only English. Participants’ views on standards of written communication in a healthcare workplace from a non-English-speaking background (NESB) was considered just as important as a native English-speaker perspective. From the total of 18 participants, four were from a NESB that included the languages Cantonese, Mandarin and Shona. Around half of the participants were born in Australia, however ten out of the 18 panel members migrated from other countries including Israel, Hong Kong, Zimbabwe, China, Sri Lanka, Singapore, Scotland, England and South Africa. Even though a range of countries of birth were represented in the study, most participants completed their medical training in Australia. Some also received their qualifications in Zimbabwe, the UK and Sri Lanka (see Appendix M).
3.6.2 Ethical clearance

A ‘Plain Language Statement’ (PLS) and ‘Consent Form’ were given to all participants (see Appendix N). The PLS informed panellists of the following: 1) what the project was about; 2) what participants would be required to do; 3) what would happen if participants agree to participate; 4) how confidentiality would be protected; and 5) what would happen if a participant no longer wished to be involved with the project at a later date. The PLS also gave information about the researchers involved with the study. The consent form asked participants to sign a declaration that they had read and understood the PLS. It also asked the researcher to again sign a declaration that a verbal explanation of the research project, its procedures and risks and had been given to and understood by the participant.

3.6.3 Instruments

Eight instruments were used with all 18 participants in the study: 1) ‘Participant Pre-Task Form’; 2) ‘Participant Background Information Form’; 3) ‘AJM Training Response Form’; 4) ‘Training Evaluation Form’; 5) Task 1 ‘Brian Edwards’ (T1-BE) and Task 1 ‘Brian Edwards’ (T1-BE) Sample Response; 6) Task 2 ‘Betty Johnson’ (T2-BJ); 7) ‘Take-home Response Form’; and 8) ‘Final Evaluation Form’ (see Appendices O, P, C, Q, A, B, D and R). Each is described in detail in this section. A further three instruments were used with the five think-aloud participants only. These were: 1) ‘Think-aloud Pre-task Form’; 2) ‘Think-aloud Response Form and Guidelines’; and 3) ‘Think-aloud Final Evaluation Form’ (see Appendices G, H and S). Each is also defined in detail in this section. A computer, digital projector, white board and white board marker were used in the training workshops. A digital recorder was used in the training workshops and think-aloud sessions.

The ‘Participant Pre-task Form’ (see Appendix O) outlined and reminded panellists what they would be required to do at the training workshop. The form also asked participants before they attended the workshop to think about and come prepared to discuss what it means for a HP to have minimally COMPETENT written English
communication skills in the workplace (i.e. in their view, what a HP ‘can do’ or should be able to demonstrate in their written communication). Also, they were asked to think about what additional aspects would make them STRONG, what may they struggle with or couldn’t demonstrate if they were NOT YET COMPETENT, and what other aspects would make their workplace written communication UNSATISFACTORY.

The ‘Participant Background Information Form’ (see Appendix P) collected demographic details from participants such as their gender, country of birth, native language, country where the majority of medical training was undertaken, profession, specialisation and current/recent workplace(s) and role(s). It also gathered information about participants’ number of years of experience since registration in Australia and whether this was in a supervisory role (if so with IMGs or not). In addition, details were collected about participants’ level of workplace interaction with NESB colleagues and approximately how many referral letters (or similar pieces of communication) written by a NESB medical practitioner they would read per month.

The ‘AJM Training Response Form’ (see Appendix C) was used by participants to make their standard-setting training judgements. The form had five spaces for panellists to note their performance level judgements. It asked participants to read a selection of five referral letters written by IMGs and if they were their supervisor, to consider each writing sample in terms of their competence to participate in entry-level/supervised clinical practice involving interaction with co-workers, supervisors and other HPs. Panellists were asked to give an overall/holistic rating of the written communication skills in each response by placing an ‘X’ in a category for each ‘writing response ID code’ and to use the ‘between’ categories as freely as the others.

Participants were asked to complete a number of evaluation forms of the standard-setting procedure and sessions. This was undertaken as part of an overall final evaluation of the standard-setting process. These were ‘Training Evaluation Form,’ ‘Final Evaluation Form’ and ‘Think-aloud Final Evaluation Form’ (see Appendices Q, R, S). These questionnaires utilized the design principles of Henn et al. (2009), Mackey and Gass (2005) and Patton (2002) which involved a mix of attitude scales and open
questions. Closed questions were not used as according to Henn et al. (2009) they may be “criticised for forcing respondents into a pre-determined response rather than letting them answer in their own words” (p. 163) and that open questions “do not force the respondent into a predetermined category that can obscure nuances” (p. 163). Regarding scaled questions, participants were asked to respond to statements and indicate their level of agreement on an attitude or ‘Likert’ scale. The scales used were either a 5-point scale ranging from ‘strongly agree’, ‘somewhat agree’, ‘neutral’, ‘somewhat disagree’, to ‘strongly disagree’ or ‘not applicable.’ Also, a 3-point scale was used ranging from ‘very influential’, ‘somewhat influential’ to ‘not at all influential’. Another scale with a range from ‘very useful’, ‘somewhat useful’ to ‘not at all useful’ was employed. These ranges were chosen, as according to Henn et al. (2009), they should be ‘balanced,’ ‘unidimensional’ and not slanted in any way so as to produce a predisposed response. The questionnaires were also partly adapted from and applied some of the work undertaken by Skorupski and Hambleton (2005) in their investigation of standard-setting panellists’ thought processes.

The ‘Training Evaluation Form’ (see Appendix Q) was completed by participants at the end of the training workshop and asked participants the degree to which they agreed with several statements. This was carried out to discover whether judges understood what they had to do and to decide if any additional training would be required. These statements included: 1) understanding of the standard-setting task; 2) whether training in the method was adequate; 3) understanding of the difference between performance levels including the ‘between’ levels; 4) understanding how to make judgements; and 5) understanding how to use the judgement form provided. Also, participants were asked whether the feedback and discussion in the training workshop would: 1) help panellists to refine their understanding of the standard-setting process; 2) help to refine their judgements and understanding of the performance levels; and 3) influence their further writing response judgements. For two judges, J1 and J10, this was not applicable as due to time and logistical constraints they could only attend a training session individually with the researcher.
The ‘Take-home Response Form’ (see Appendix D) was nearly identical to the ‘Training Response Form’ (see Appendix C). The only difference was that the take-home form had 30 spaces for participants to make judgements about writing responses, whereas the training form had five spaces. All participants received a ‘Take-home Pack’ that included a ‘Take-home Response Form’, 30 test candidate writing responses to judge (see section 3.6.4 for how the responses were allocated to participants) and a ‘Final Evaluation Form’. The ‘Final Evaluation Form’ asked participants how influential a number of statements were when making judgements. This was undertaken as an evaluation of the standard-setting process overall. The statements included: 1) participants’ experience with the type of writing task from their working life; 2) participants’ sense of what test candidates should be able to do in order to be proficient/safe in the workplace; 3) participants’ own definitions of performance levels; 4) other participants’ definitions of performance levels; 5) feedback/comments/discussion with other participants; and 6) writing response judgements/ratings of other participants. Panellists were also asked how useful a number of statements were in judging writing responses. The statements included: 1) practicing the procedure; 2) feedback/comments/discussion with other participants; and 3) writing response judgements of other participants. Once again, this was not pertinent for two panellists (J1 and J10) for some of the statements, as they could only attend a training session on their own with the researcher.

The ‘Think-aloud Pre-task Form’ (see Appendix G) outlined to participants what would be involved during the think-aloud session. This included informing participants that as they read and judged each script they would be asked to say all their thoughts ‘out loud’ on the spur-of-the-moment and unprompted. They were also asked to include comments on what aspects captured their attention and the factors they took into consideration in deciding on a performance level. The pre-task form directed participants to view a ‘think-aloud’ example of a doctor testing a new web page’s usability: https://www.youtube.com/watch?v=nJ2udLjdsx4. The ‘Think-aloud Response Form’ (see Appendix H) was again similar to the ‘Training Response Form’ and ‘Take-home Response Form’ (see Appendices C and D). A difference was that the think-aloud form had 12 spaces for participants to make judgements about writing
responses. It also reminded participants about the AJM standard-setting method and what was involved during the think-aloud session (i.e. for participants to say their thoughts ‘out loud’ as they judged writing samples).

The ‘Think-aloud Final Evaluation Form’ (see Appendix S) was completed by participants at the end of the think-aloud session. This was again completed as a final evaluation of the think-aloud stage of the overall standard-setting process. The ‘Think-aloud Final Evaluation Form’ asked participants the degree to which they agreed with several statements. These statements included whether: 1) the instructions were clear and participants understood what they were required to do when ‘thinking aloud’; 2) the level of prompting from the researcher was adequate; 3) participants were comfortable with the number of writing responses judged and spoken about; 4) participants had adequate time to speak comprehensively about each writing response; and 5) the process of ‘thinking aloud’ didn’t alter the way in which participants judged the writing responses. Participants were also asked about any additional comments concerning the experience of ‘thinking aloud’ overall.

Two hundred writing samples for the standard-setting sessions were provided by Cambridge Boxhill Language Assessment. These were scanned copies of original test candidate writing responses from retired tests. The test candidates’ writing scripts were from a variety of first language backgrounds. The writing samples had been written in response to two tasks: Task 1 ‘Brian Edwards’ (T1-BE) and Task 2 ‘Betty Johnson’ (T2-BJ) (see Appendices A and B). Both tasks require test candidates to write a letter of referral. All OET Writing sub-test tasks are designed and considered to be of equal difficulty for test candidates. The writing responses had a clear identifier on the top right corner (e.g., Task 1 = T1-BE-89 and Task 2 = T2-BJ-35). One hundred writing samples for each of the two tasks were used (with an associated fair average score, generated by a many-facet Rasch analysis, based on the ratings of 2-3 English-language trained professionals from Cambridge Boxhill Language Assessment). In producing a fair average score, the Rasch analysis uses FACETS software (Linacre 2017) to account for and counterweigh OET raters whose scores are especially severe or
lenient. The writing samples for both tasks included a range of OET band scores (A-D) or approximate AJM performance level categories (STRONG-UNSATISFACTORY) and associated fair average scores. At present, the lowest ranking used for OET results is grade E. This is, however, seldom awarded in usual test administrations and hence grade E was omitted for the purposes of this study. The sample writing response for Task 1 ‘Brian Edwards’ (T1-BE) (see Appendix A) was conceived and written in consultation with a HP by a language professional. It is intended as a sample response only and not a ‘perfect’ example.

3.6.4 Procedures

The procedures for the main study are outlined in this section. These include participant recruitment, writing response distribution and allocation to participants, workshop procedures and data analysis procedures (qualitative and quantitative).

3.6.4.1 Participant recruitment

Participants were recruited using a variety of methods. Initially, purposive sampling was used. Health professional members from the ARC interdisciplinary project group (mentioned in introduction section 1.5) were requested to contact their medical practitioner colleagues to generate potential interest and involvement. Some of these contacts had previously participated in the earlier spoken communication phase of the overall ARC linkage project. From this initial list of contacts, snowball or chain sampling was utilized. Panellists who agreed to participate were asked to contact their professional networks and generate further involvement in the project. Also, approximately one hundred ‘cold call’ emails and letters were sent to medical professionals in Melbourne. Another participant was recruited through personal contacts of the researcher. One was also enlisted via a University of Melbourne healthcare communication student reading group. Participants were offered compensation for their time.
In the initial recruitment phase, participants were sent the ‘Plain Language Statement’ (PLS) (see Appendix N) and general information about the study. After this, participants who agreed to take part were invited to attend a workshop at a convenient time and date. The further general information outlined that throughout the study they would be asked to give judgements about writing responses from the OET’s Writing sub-test and that these were written by IMGs whose first language is not English. Panellists were also sent a ‘Consent Form’ (see Appendix N).

In addition, the participants were given a pre-training workshop task (see Appendix O). This task informed panellists that in the upcoming workshops the descriptions that would be used for each performance level standard: STRONG, COMPETENT, NOT YET COMPETENT and UNSATISFACTORY. Before attending the workshop, participants were asked to come prepared to discuss what it means for a NESB overseas-trained HP to have minimally competent written English communication skills in the workplace. In other words, in their view, what should a HP be able to demonstrate or ‘can do’ in their written communication to be COMPETENT. Furthermore, what additional aspects would make them STRONG, what may they struggle with or couldn’t demonstrate if they were NOT YET COMPETENT and what other aspects would make their workplace written communication UNSATISFACTORY.

Participants were also sent a copy of writing Task 1 ‘Brian Edwards’ case notes and a sample response. This example response was not written by an OET test candidate, but rather a task developer (language expert) from Cambridge Boxhill Language Assessment with content guidance from a health professional (domain expert) – see Appendix A. The writing task and sample response was sent ahead of the training workshop in order to allow more time for discussion about participants’ views on levels of language proficiency and comparisons of performance level judgements. The training workshop and think-aloud session procedures are outlined in section 3.6.4.4.
3.6.4.2 Writing response distribution

As mentioned, one hundred writing samples for each of the two tasks (200 total) were used: Task 1 ‘Brian Edwards’ (T1-BE) and Task 2 ‘Betty Johnson’ (T2-BJ) (see Appendices A and B). In distributing these writing responses to the 18 participants, a number of issues were considered. These were that there should be a balance between: 1) overlap of writing samples judged by participants in order to check rating consistency; and 2) a maximum range of judgements (across the whole 100 writing samples per task) to maximise the number of scripts that were rated overall. The more responses that were rated meant the greater the probability that there would be a reasonable number of performance level allocations in each category (including the ‘in-between’ categories which are crucial for this method). In order to achieve this, a minimum of three judges were allocated to consider the same writing responses with some overlap for consistency and reliability checking. In most cases, a group of three or four participants judged two, three or four scripts that overlapped. However, in order to realize a maximum spread over the full range of 100 samples per task this was sometimes slightly more or less. Participants who completed stages 1-2 (training and take-home allocation) were asked to consider 35 writing responses overall (five training workshop responses and a further 30 samples to take home and judge). The five ‘think-aloud’ participants (stage 3) judged another 12 scripts (47 overall).

3.6.4.3 Writing response allocation to participants

The writing responses were allocated to participants in the following way. For the training workshops, participants used the same core of five writing samples from Task 1 ‘Brian Edwards’ (T1-BE). This was undertaken to assess participants’ rater consistency across this task. All participants started with sample T1-BE-74 (OET band score C or NOT YER COMPETENT, OET fair score 4.77) (see Appendix T). This script was chosen because it was close to the current OET B/C OR AJM COMPETENT/NOT YET COMPETENT cut score divide of 4.8 (rather than in the middle of OET band B or C, AJM COMPETENT/NOT YET COMPETENT) and that this might stimulate initial discussion. At this early point it was not known whether the current cut scores were at all what
health professionals considered appropriate or not, but it was thought to be the best starting point for participant debate. The remaining four Task 1 ‘Brian Edwards’ (T1-BE) writing responses were presented in random order for each workshop. They were allocated randomly to alleviate any potential ‘order effect’ (Sakyi, 2000) where a participant’s judgement of writing samples may be influenced by what has previously been read (Milanovic et al., 1996).

In stage two, panellists received a take-home pack of 30 new test candidate responses to judge. This included 12 new responses from the Task 1 ‘Brian Edwards’ (T1-BE) training workshop task and 18 responses from the new Task 2 ‘Betty Johnson’ (T2-BJ) task. There was again a common core of five Task 2 writing responses in order to check participants’ rating consistency across this task. These samples were T2-BJ-2, 4, 19, 20, 22 that included a range of previously rated OET band scores (A-D) or AJM (STRONG-UNSATISFACTORY) and raw scores (Cambridge Boxhill Language Assessment associated fair average scores, generated by a RASCH analysis, based on the ratings by 2-3 English-language trained professionals). Once more to alleviate an order effect, all of the 30 writing responses (including the common core of five Task 2 ‘Betty Johnson’ (T2-BJ) samples) in the participants’ take home packs were arranged entirely at random to make it less likely that they might make task-specific decisions based on their perception of how difficult the task was. This meant that participants may have read, for example, two Task 1 responses, three Task 2’s, then one Task 1 and another four Task 2’s. It was thought that by randomizing the writing response order and tasks this would make it more likely that participants would focus on the performances themselves and would remove any potential task factors from consideration.

In stage three, as part of a further qualitative investigation, five interested participants (optional participation only) were asked to judge an additional 12 OET Writing responses and record their thought processes while rating these scripts via a think-aloud protocol (TAP). As for the allocation of writing responses to the five ‘think-aloud’ participants, it was deemed important that the judges be allocated scripts that they had not already rated. This is so they would be coming to the set of samples afresh, but with the background of having rated other samples. The five think-aloud
participants were Judges 6, 9, 12, 15 and 18. These judges were allocated writing scripts in this way to achieve maximum spread over the 100 rating samples and to ensure that these participants had not previously rated the TAP writing responses. In addition, this allocation allowed the 12 TAP writing samples to have been read and judged by other stage two panellists.

The 12 think-aloud writing responses were chosen from ‘Brian Edwards’ (T1-BE). These samples again covered a range of previously rated OET grades (A-D) and scores (Cambridge Boxhill Language Assessment associated fair average scores, generated by a many-facet Rasch analysis, based on the ratings by 2-3 English-language trained professionals). During the think-aloud session, to again avoid any order effect, these 12 scripts were presented randomly to each of the five participants. The 12 Task 1 ‘Brian Edwards’ (T1-BE) writing responses had also been allocated to five other participants (Judges 1-5) to rate as part of the stage two standard-setting procedure.

3.6.4.4 Workshop procedures

The workshop procedures are outlined in the following sections. All 18 participants completed stage one which involved a workshop training session. However, due to time constraints and logistical considerations it was not possible for all panellists to convene at the same time and date. Therefore, seven training workshops in total were conducted which included one, two, three and six participants in each session.

3.6.4.1 Training workshops

In the training workshops, participants were initially greeted and asked to fill in the ‘Consent Form’ and ‘Participant Background Information’ form (see Appendices N and P). Background about the project/OET and an overview of standard setting was delivered via a PowerPoint presentation. Participants were referred to the ‘Pre-task Form’ (see Appendix O) and invited to discuss and compare their definitions of the performance levels: STRONG, COMPETENT, NOT YET COMPETENT and UNSATISFACTORY. Panellists then read one referral letter (Task 1, ‘Brian Edwards’,
sample number T1-BE-74) (see Appendix T) written by an IMG and used the ‘Training Response’ form (see Appendix C) to judge the script and allocate it to a performance level. Participants were advised to give an overall/holistic rating of the written communication skills in each response and to exclude from their decisions test-candidates’ clinical knowledge (if evident) in the response. Panellists shared their decisions, discussed any differences and specified why they placed the response in a specific performance level category. If there were any areas of substantial disagreement, they were asked to discuss why. Participants were free to change their placement of a writing response on their own individual form after receiving feedback from other participants, but they did not have to do so (two of the training sessions only involved one participant and therefore this was not applicable). Participants then read a further selection of four referral letters allocated from Task 1, ‘Brian Edwards’ that were presented in random order for each workshop and completed the same as above (i.e. shared their decisions and discussed differences, if applicable). A digital recorder was used to record the standard-setting workshop discussions. Panellists completed and returned the ‘Training Evaluation’ form (see Appendix Q). A take-home pack that included 30 new test candidate responses, ‘Take-home Response Form’ and ‘Final Evaluation’ form (Stage 2) (see Appendices D and R) was distributed. Participants were informed of stage three which was a further qualitative section of the study. Interested participants were emailed to arrange a convenient time/location for the think-aloud protocol (TAP) session. Panellists were thanked for their participation and the session was concluded. Post-training, materials were collected for transcription/coding.

3.6.4.2 TAP sessions

Panellists who agreed to be further involved in the stage three think-aloud protocol (TAP) section of the study (five total – J6, J9, J12, J15, J18) were sent the ‘Think-aloud Pre-task Form’ (see Appendix G). This form gave general information/instructions about what participants would be required to do during the session and a think-aloud example. At the pre-arranged time and place, participants were reminded in written and verbal form what they would be required to do during the session using the
‘Think-aloud Pre-task Form’. This think-aloud information sheet provided a guide to the overall procedure and applied the developmental work of Cohen (2000) and Ericsson and Simon (1993). These instructions were also repeated to participants orally. A digital recorder was used to record the think-aloud sessions. Participants were instructed to give a holistic judgement of the writing proficiency of the test candidate in each script and to concurrently verbalise their thoughts ‘out loud’ while forming these judgements. Each participant was asked if had they had any queries about what was required of them during the TAP session and if they understood the instructions.

Panellists were then reintroduced to the task case notes – Task 1 ‘Brian Edwards’. This task was used during the stage one training session and part of the stage two take-home evaluation pack. The Analytic Judgement method (AJM) standard-setting procedure was again used in the main study’s think-aloud sessions (as in the trial) and participants were reminded of the method and the form to be used – a variation of the training workshop and take-home forms (‘Think-aloud Response Form and Guidelines’ – see Appendix H). 12 writing responses were presented in random order as in the think-aloud trial (see section 3.5.3.4). Also, as in the trial think-aloud sessions, the researcher’s prompts were ‘non-mediated’ (Ericsson & Simon, 1993; Green, 1998).

Once participants had spoken out loud about the 12 writing responses, they were asked to complete a short, structured interview about the process and procedure of thinking aloud. These questions were similar to those stated on ‘Think-aloud Evaluation Form’ (see Appendix Q). In these post-interviews, participants gave expansive and elaborated responses to the process of thinking-aloud. Finally, participants completed the ‘Think-aloud Evaluation Form’ (see Appendix S) and indicated their level of agreement with a number of statements on a 5-point ‘Likert’ scale. The think-aloud sessions (including post interviews) took approximately 60 minutes in total and were carried out individually.
3.6.4.5 Data analysis procedures

The following section outlines the data analysis procedures used in the main study. The qualitative analysis procedures are shown first and then the quantitative procedures.

3.6.4.5.1 Qualitative analysis procedures

The panel members’ training workshop discussions and think-aloud protocols (TAPs) were recorded. These recordings were then transcribed. Repetitions, false starts and pauses such as ‘ah’ or ‘um’ were not included as it was felt that these did not contribute anything meaningful to the content analysis of their comments, which was the focus of the investigation. Short pauses of one to two seconds were shown with line breaks. Full stops and commas were used to show ‘brief’ and ‘very brief’ pauses. Longer pauses of more than two seconds were indicated, but overall these were minimal (apart from some participants who read silently to themselves during short sections of the TAP sessions and training workshops). Intonation, tone and emphasis were not indicated because as Someren, Barnard and Sandberg (1994) point out, this kind of data transferred into written form could be seen as extremely interpretive and therefore unreliable and also was not useful for the study’s purpose.

The following modes of commentary were used for the transcribed workshop discussions and TAPs: participant verbalization was shown in plain text; participants reading aloud from a test candidate referral letter or task case notes (verbatim or paraphrase) was indicated in italics; researcher talk was shown in UPPER CASE; single quotation marks were used for direct speech reported by participants; double quotation marks were used for participants’ performance level judgements; incomplete words or interrupted restarts within a word were indicated with a hyphen (-); speakers’ incomplete utterances or interrupted utterances (where final word spoken is complete) were shown with ellipses (...); round brackets were used for possible interpretations of unclear syllable(s)/word(s); and square brackets were used for the researcher’s clarification of participant utterances that were unclear.
Once the data had been transcribed and prepared, a thematic analysis of the participants’ comments and value judgements was subsequently conducted. The intention of the thematic analysis was to investigate the research questions: 1) What criteria do participants use as a basis for their decisions in judging writing responses and to what extent are these decisions language based?; and 2) Is there any variability between judges in what they attend to while setting standards? The overarching themes and more specific categories/codes for the thematic analysis were derived inductively from the data themselves based on Glaser and Strauss’s (1967) formulation of ‘grounded theory’ in which the ‘naïve’ role of the coder is implied (Strauss & Corbin, 1990; Vaus, 2002). Thus, as Hycner (1985) states, in grounded theory methodology, the researcher tries to “stay as true to the data as possible” (p. 282) and the development of themes and codes is ‘data driven’ as it is the data that largely decides what the categories will be rather than any pre-determined ones instituted by the researcher. However, it may be necessary to employ a modified grounded theory to acknowledge that a ‘tabula rasa’ or completely blank slate may not be feasibly achieved in arriving at themes because pre-conceived standpoints may be prevalent (Vaus, 2002). Patton (2002) proposes two phases to analyse qualitative data:

First, the analyst can identify, define, and elucidate the categories developed and articulated by the people studied to focus analysis. Second, the analyst may also become aware of categories or patterns for which the people studied did not have labels or terms, and the analyst develops terms to describe these inductively generated categories (p. 454).

In a grounded theory approach, a common method of code development and assigning of data to these codes is a recursive process of ‘constant comparison’ (Miles & Huberman, 1994). In this process, initial codes are identified following a preliminary scan of the data or sub-set of the data. Then these initial codes are subsequently updated and refined after additional scans and this recursive process is further carried out until the codes satisfactorily encapsulate the whole dataset. This process was used to develop codes for this study.
3.6.4.5.1.1 Applying codes

When applying codes there are a number of considerations that need to be taken into account. Each of these is examined in brief to back the approach taken in the current study. In applied linguistics, a common method when coding qualitative data is to beforehand divide the data into segments (Green, 1998) in order to, for example, count them. Pieces of data may merely be an individual word/phrase or extended section of text (Punch, 1998). Some studies, such as in second language acquisition, may seek to quantify errors or number of words in each segment and this in itself is an area of research (e.g., Foster, Tonkyn & Wigglesworth, 2000). However, in thematic analysis the subject of enquiry is on the actual content of the data instead of the structure and this type of segmentation may hinder the investigation. Therefore, the prior segmentation of pieces of data was not undertaken in the present study as the focus was on the “informational content” (Forman & Damschroder, 2008, p. 40), instead of how language is actually used. Furthermore, there was no need to quantify the data because even if a particular code/category was referenced more frequently it was not assumed that this indicated greater significance relative to other codes. Rather, it was considered more important that codes/categories were appropriately validated (see section 3.6.4.5.1.2 below) and applied consistently by the researcher and any overlap was noted for additional enquiry.

When coding data, an issue that may arise is that more than one code or category might feasibly be applied to a section/s of the dataset. Also, precisely where the division should practically be made between words or phrases in order to accurately correspond to each code/category may prove problematic. A solution to this issue is to allow more than one code/category to be applied if required. Brice (2005), in an analysis of ESL students’ semi-structured interviews, highlighted this issue of coding data when more than one possible code/category might apply. She observed that “the remarks may be organized in such a way as to make it impossible to tease them apart without losing context necessary for interpretation of those belonging to either category” (Brice, 2005, p. 164). Similarly, in the present study, it was practical to retain the data as intact segments rather than divide them into analytic units before coding.
because participants’ comments often expanded over a long stretch of talk with several ideas embedded within it. Also, it was considered that if the data was de-contextualised too much this would result in a possible loss of meaning which might further weaken the accuracy of the coding process. Brice (2005) encountered issues in attempting to “unitize and then categorize” interview transcripts (p.162). She found that boundaries were too concentrated and hence the meaning was too de-contextualised. This led her to use coding categories and unit boundary divisions as ‘episodic units’ (Grant-Davie, 1992) in which they last “for as long as a participant continues to make the same kind of comment” (p.163). The current study took this latter approach when dealing with issues of coding boundaries and category overlap.

The researcher, through a process of recursive data examination and re-examination, refined the codes and categories to produce a ‘Data Dictionary’ (see Appendix U). The first phase was a ‘bottom up’ process based purely on the domain experts’ insights and the second was a process of organizing themes into overarching or general categories. The data dictionary included several general themes: ‘Task Fulfilment’ (TF), ‘Content’ (C), ‘Organisation’ (O), ‘Expression’ (E), ‘Presentation’ (PR), ‘Professionalism’ (PRO) and ‘Other’ (OTH). Each of these was further divided into specific categories or codes. The same data dictionary was used to code both the standard-setting workshop discussions (participants’ performance level definitions and judgements) and think-aloud protocol (TAP) sessions. Before being implemented, the data dictionary was revised several times. This included feedback from applied linguists at the University of Melbourne who read a sample from a training workshop and think-aloud transcript. Advice and suggestions were given to hone the data dictionary for it to be further used in the coding validation process with an ‘official’ second coder. Also, a former participant (J15) in one of the main workshop discussions and think-aloud sessions gave additional input to refine the data dictionary. Perspective and input from a medical expert in developing the codes was valuable as the researcher does not have a health background. A sample of a training workshop and think-aloud transcript, already coded by the researcher, was given to participant J15. She was asked to read through the transcripts and confirm whether the preliminary codes/categories were effective or not and were being applied consistently.
J15 made several suggestions to improve the initial data dictionary. She commented that there may be overlap between items (e.g., ‘General Organisation’ (GO) and ‘Discourse Structure’ (DS), which perhaps could be combined). It was decided to pool the two codes ‘General Organisation’ (GO), ‘Discourse Structure’ (DS), but to maintain ‘Prioritizing’ information (PR) as separate due to subtle differences. ‘General Organisation’ (GO) was initially used for more general comments about the overall organization of a test candidate’s letter, however it was found that this could be encapsulated in the code ‘Discourse Structure’ (DS) which was related to more specific comments about the sequence of ideas, stages, tasks, actions and processes in the letter. Furthermore, the code ‘Prioritizing’ (PR) key details/information for the reader was kept distinct as it more specifically related to the test candidate putting more important information ‘up front’, such as in the first paragraph, before less important information.

J15 also suggested that the code ‘Current Medication’ (CM) could be split into two codes: the current new medication (e.g., flucloxacillin/antibiotics) and the usual regular medications. J15 stated: “the (CM) code tended to encompass both the antibiotics and usual regular medications, but from the transcript, it was often only referring to the antibiotics”. Therefore, a new code for ‘Regular Medications’ (RM) was created. Additionally, J15 pointed out that a “diagnosis/provisional diagnosis code” was needed as “this was not really covered by other codes such as ‘Clinical Examination’ (CE)”. Therefore, a new code for ‘Diagnosis/provisional diagnosis’ (D) was created. Finally, J15 pointed out that “sometimes allergies, family history and smoking were mentioned” and considered whether these should be combined into an overall ‘Content’ category or to code them separately. It was decided not to create separate codes/categories for these, but to include allergies in the ‘Medical History’ (MH) code and smoking and family history in the ‘Social History’ (SH) category. It was felt that these items could be adequately included within these greater codes/categories (see ‘Data Dictionary’ – Appendix U). The following section discusses how the study’s coding procedure was validated.
3.6.4.5.1.2 Coding validation

The data dictionary was validated using a second coder who had no previous involvement with the study. This coder had previously worked as an English as a Second Language (ESL) teacher and also had a Masters in Applied Linguistics. The second coder was trained by the researcher to use the data dictionary and apply the codes to a small initial section of a sample transcript. Together, the second coder and researcher coded this sample until both felt confident in applying the codes/categories successfully and consistently. This involved assigning codes/categories to sections of the transcribed data. Also, the double coder was asked to indicate or double-code any overlap with other codes/categories or segments of the data. During this sample coding session, each coder identified a similar number of segments with only a slight degree of disagreement. Disagreements concerning segmentation were mainly due to the classification of each new referent in the transcript (i.e. the section of the transcript that was being referred to such as a single word, sentence or phrase). There was also disagreement over how to separate and code repetitions of participants’ comments. These coding disagreements were settled by revising the data dictionary to contain more detailed instances of the individual codes and their uses. This included examples of code overlap in an utterance (if applicable) and situations where a code may also relate to another code. An example of this is shown in table 3.1 below. In this utterance, the code ‘Prioritizing key details/information for the reader’ (PR) from the theme ‘Organisation’ overlaps with the code ‘Relevant past medical history evident/or not’ (MH) from the theme ‘Content’. The overlapping coded section is shown in bold. As mentioned, further instances of code/theme overlap can be seen in the full data dictionary (see Appendix U).
<table>
<thead>
<tr>
<th>Theme/ Code</th>
<th>Category Description</th>
<th>Example</th>
<th>Code overlap in an utterance (if applicable – in bold)</th>
<th>Also relates to code (if applicable)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PR</td>
<td>Prioritizing key details/information for the reader e.g., putting more important information before less important information for the reader</td>
<td>Even if I was going to write it like this, I'd still put, I'd put the SCC first and then say, ‘Other than that they have glaucoma.’</td>
<td>MH</td>
<td>MH</td>
</tr>
</tbody>
</table>

### Table 3.1: Code/theme overlap

The second coder pointed out some initial issues with the data dictionary and these are discussed and addressed as follows. The second coder noted that under the theme ‘Content,’ the code ‘Regular medication listed/or not’ (RM) was slightly confusing and similar to the code ‘Relevant past medical history evident/or not’ (MH) (i.e. diseases/symptoms/conditions/allergies to the presenting complaint). While it was acknowledged that the (RM) code did overlap to some extent with a patient’s past medical history, it was decided to keep the (RM) code as it related to specific examples of medications that were mentioned/or not by the test candidate. The domain expert second coder (J15) who initially helped to refine the data dictionary expressed the view that it was important to differentiate these two codes and hence it was decided to keep them as two separate codes.

A further concern under the theme of ‘Presentation’ was with the initial code ‘Overall appropriateness/or not of letter layout/format’ (LO). That is, general comments about: 1) the use of standard formal letter layout (e.g., use of date, writer/sender address) and where these are situated on the page; and 2) text spacing (e.g., writing on every/second line)). The researcher and second coder noted that this code was too dense and included too many specific details to encompass a single code. It was decided to split this (LO) code into parts and create new codes. The code ‘Overall appropriateness/or not of letter layout/format’ (LO) (i.e. general comments about the use of standard formal letter layout) was retained to refer to overall general comments about the letter’s layout. A new code ‘Use of date/or not’ (DA) was created...
to refer to specific references to a date being included/or not in the letter. This code was moved from the overall theme of ‘Presentation’ to ‘Content’. In addition, ‘Writer/sender address evident/or not and where this is situated on the page’ (WA) was collapsed and merged with the ‘Doctor identification mentioned generally as evident/or not’ (DID) code (e.g., writer/sender’s title, address, contact details, department) to indicate whether the test candidate had included their own address for the reader (or not) under the theme of ‘Content’. A final new code ‘Text spacing’ (TS) (i.e. writing on every/second line) was created to refer to the spacing of text on the page.

After these issues with the data dictionary had been addressed and revised in this way, these instances of coding disagreement were resolved. The second coder then independently coded a further 20% sample of the dataset using the updated data dictionary. This same 20% sample was also coded separately by the researcher. Inter-coder reliability checks were carried out to ascertain levels of agreement between the two coders (second coder and researcher). For this purpose, two suggested measures were utilized (Lombard, Snyder-Duch & Bracken, 2010; Someren et al., 1994): percent agreement and Cohen’s kappa (k). Lombard et al. (2010) argue that even though percent agreement is highly used and is quite easy to calculate, the literature on its usage is inconsistent and it may be regarded as a deceptive measure that may misjudge and overvalue true inter-coder agreement. A disadvantage of the percent agreement measure is that it does not remove chance agreements from the overall proportion. This is likely to occur the more frequently a code is used. In the current study, some codes were used more frequently than others. Hence, an additional measure, Cohen’s kappa (k), was also used to take into account chance agreements. Kappa contrasts the actual number of coding agreements with the number of agreements that would be anticipated due to chance. Lombard et al. (2010) state that, like many inter-coder reliability measures, Cohen's kappa has its own advantages and disadvantages, but that it should still be considered the measure of choice.

The reporting of the two inter-coder reliability measures was calculated using ReCal software (Freelon, 2010). Percent agreement between the researcher and the second
coder was 93%. This indicates an acceptable level of agreement. The overall Kappa value was 0.91 which again shows a tolerable level of inter-coder agreement. The literature points to differing levels of acceptability for an ‘excellent’ or ‘high’ level of Kappa agreement. Mackey and Gass (2005) state that a range of Kappa values from 0.81 - 1.00 would suggest an ‘excellent’ agreement level. Furthermore, in a think aloud study with essay raters, Wolfe et al. (1998) state that agreement values of 0.85 – 0.93 are ‘acceptable.’ Therefore, this study’s Kappa value of 0.91 would suggest that it is adequate. Having determined an acceptable level of inter-coder reliability for the coding categories and also having the data dictionary validated, the researcher further coded the participants’ training workshop discussion and think-aloud session transcripts. This was undertaken with the assistance of QSR International’s (2014) NVivo 11 qualitative research software. The main reason for using NVivo software was its relative ease for categorising, labelling and demarcating units of analysis. It was also chosen for its ability to offer an easily accessible place for storing, retrieving and organising data that assisted in the analysis. The results of the qualitative analysis are shown in the results section 4.2.

3.6.4.5.1.3 Standard-setting evaluation

All participants were asked to complete two overall evaluation forms of the standard-setting procedure: ‘Training Evaluation Form’ and ‘Final Evaluation Form’ (see Appendices Q and R). Five of the 18 panellists who participated in the TAP section of the study also completed a ‘Think-aloud Evaluation Form’ (see Appendix S). Participants were asked whether they agreed or disagreed with a number of statements on a 3-point or a 5-point ‘Likert’ scale (see section 3.6.4.4.2). The participants’ responses to these statements were combined and totalled for each point on the scale. In addition, participants were asked to express any further feedback or comments regarding the standard-setting procedure. These comments were analysed and ordered thematically. The results of the training evaluation, final evaluation and think-aloud evaluation are shown in the results section 4.3.
3.6.4.5.2 Quantitative analysis procedures

Firstly, in this section, a description of how cut scores were calculated using the Analytic Judgement method (AJM) is stated. Then, an account of how participants’ rating behaviour was attended to in the study is given.

3.6.4.5.2.1 Calculation of cut scores

Cut scores between the performance levels STRONG, COMPETENT, NOT YET COMPETENT and UNSATISFACTORY were established post-workshop using the tallied results of the ‘between’ performance level categories of participants’ judgements. The Analytic Judgement method (AJM), as described previously, was used to calculate the cut scores. The scores utilized for this calculation were the fair scores provided by Cambridge Boxhill Language Assessment. As noted, these fair scores were determined by a previous FACETS (Linacre, 2017) analysis of the OET raters’ raw scores which takes into consideration the relative severity of each language professional rater.

The workshop data were entered into a Microsoft ‘Excel’ spreadsheet. All the fair scores previously assigned to the writing responses that were placed by the domain expert judges in the ‘between categories’ (e.g., between STRONG and COMPETENT) were added together. If a particular writing script had been placed in a between category by two or more judges, the fair score previously assigned to it was counted twice, three times etc. (as applicable) in the overall sum. The total of all the combined fair scores for each between category was averaged to produce a mean score. The mean of the fair scores for the performances in the between categories is the new cut score. The cut scores derived from the OET fair scores were represented on a scale from 1 to 6. New cut scores were determined for the categories ‘between STRONG and COMPETENT’, ‘between COMPETENT and NOT YET COMPETENT’ and ‘between NOT YET COMPETENT and UNSATISFACTORY’. AJM performance levels correspond to these OET bands: STRONG = A; COMPETENT = B; NOT YET COMPETENT = C; and UNSATISFACTORY = D. The results of these calculations are shown in the results section 4.4.
3.6.4.5.2.2 Judge behaviour

An analysis of the standard-setting panellists’ judgements of writing scripts was carried out using the software program FACETS version 3.8.0 (Linacre, 2017). FACETS implements a many-facet Rasch model which is an extension of the first basic Rasch model developed by Georg Rasch in 1960. McNamara (1996) argues that the many-facet Rasch approach is a valuable model for undertaking quantitative analyses as “much lies beneath the surface in performance testing, and that Rasch analysis is useful in revealing underlying patterns in ratings data which can be interpreted in ways that raise fundamental questions of test validity” (p.216). Rasch’s (1960) initial mathematical depiction showed the interaction between item difficulty and test taker ability on an interval scale of ‘logits’ (McNamara, 1996). This model was first used for dichotomously scored results such as correct/incorrect scoring. It was developed further by Andrich (1978) into a rating scale model and by Masters (1982) and Wright and Masters (1982) into a partial credit scoring approach. The multi-faceted Rasch model used in this study’s analysis, as advanced by Linacre (2017), further extended the partial credit and rating scale models. Linacre’s model was able to deal with any number of facets relevant to the test context and typically included elements such as ‘test candidates’, ‘raters’ and ‘items’ as facets (McNamara, 1996). The analysis can indicate how individual elements are functioning such as the relative leniency/harshness of raters, test candidate ability and prompt difficulty and is expressed on a logit scale or log-linear metric (Bond & Fox, 2015).

This study’s data specification used the following facets for analysis: ‘candidate’ (200 elements); ‘rater’ (18 elements) and ‘task’ (2 elements). It is part of the OET’s test construct and design that all writing tasks are intended to be of similar difficulty. This is in order for the resulting test scores to be fair for all test candidates. Based on this assumption, the two selected OET Writing tasks were treated as being of equal difficulty, and therefore anchored at the same value (0) to avoid the analysis resulting in disjoint subsets. This would have happened because test candidates either responded to one of the writing tasks or the other, but not to both. For the purpose of the analysis, the AJM performance levels were recoded into a rating scale. The AJM
performance levels: ‘STRONG’, ‘between STRONG and COMPETENT’, ‘COMPETENT’, ‘between COMPETENT and NOT YET COMPETENT’, ‘NOT YET COMPETENT’, ‘between NOT YET COMPETENT and UNSATISFACTORY’ and ‘UNSATISFACTORY’ were converted to a 7-point scale from 7 – 1 with seven corresponding to STRONG and one to UNSATISFACTORY. The results of the Rasch analyses are illustrated in the results section 4.4 and responds in part to research question two: Is there any variability between judges in what they attend to while setting standards?

3.7 Argument-based validity framework and sources of evidence from this study

As noted in literature review section 2.6.4, an argument-based approach to validity utilised in this study is relevant to the ‘decisions’ inference as proposed by Knoch and Macqueen (in preparation). In their framework for an LSP test, a number of potential sources of backing were stated that could be gathered in relation to the associated warrant and relevant assumptions. The following table 3.2 highlights the ‘actual’ sources of evidence that is sought in order to provide suitable backing for validity claims for the present study.

<table>
<thead>
<tr>
<th>Warrant</th>
<th>Assumptions</th>
<th>Sources of backing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimates resulting from the performance on the LSP tasks are useful for decision-making about readiness (or similar purposes) for work in the TLU domain</td>
<td>The standards set on the LSP test reflect the language standards operating in the target language use (TLU) domain&lt;br&gt;&lt;br&gt;Relevant to research question 1 (i.e. What criteria do standard-setting participants use as a basis for their decisions in judging writing responses and to what extent are these decisions language-based?)</td>
<td>Involvement of domain experts in standard setting&lt;br&gt;&lt;br&gt;In this study:&lt;br&gt;- Qualitative comments provided by 18 domain expert participants from a range of health professions: general practitioners (GPs), specialists, consultants and medical educators</td>
</tr>
<tr>
<td>Standard-setting panellists are oriented to construct-relevant features of the performance</td>
<td>Verbal protocols during standard setting; discussions during workshops&lt;br&gt;&lt;br&gt;In this study:</td>
<td></td>
</tr>
</tbody>
</table>

**Decisions inference**: assumes that the decisions made based on the LSP test are appropriate and equitable

**Claim**: decisions made based on the estimates of the quality of the performance are appropriate and well communicated
<table>
<thead>
<tr>
<th>Research Question</th>
<th>Methodology</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. How do standard-setting panellists use as a basis for their decisions in judging writing responses and to what extent are these decisions language-based?)</td>
<td>Qualitative data from 7 training workshops and 5 verbal protocols (think-alouds – TAPs) sessions</td>
</tr>
<tr>
<td>2. Is there any variability between judges in what they attend to while setting standards?</td>
<td>Statistical analysis of score data; analysis of verbal protocol/discussion data</td>
</tr>
<tr>
<td>Relevant to research question 2 (i.e. Is there any variability between judges in what they attend to while setting standards?)</td>
<td>In this study:</td>
</tr>
<tr>
<td></td>
<td>- Analysis of AIM cut score data and subsequent use of many-facet Rasch program FACETS (Linacre, 2017)</td>
</tr>
<tr>
<td></td>
<td>- Thematic analysis of verbal protocol/discussion data using NVivo software</td>
</tr>
<tr>
<td>3. How do standard-setting judges view the process and outcome of the standard-setting procedure?</td>
<td>Interviews, surveys with standard-setting panellists</td>
</tr>
<tr>
<td></td>
<td>- Qualitative commentary from 7 training workshops and 5 verbal protocols (think-alouds – TAPs) sessions</td>
</tr>
<tr>
<td>Relevant to research question 3 (i.e. How do standard-setting judges view the process and outcome of the standard-setting procedure?)</td>
<td>In this study:</td>
</tr>
<tr>
<td></td>
<td>- All 18 participants completed a training and final evaluation form (see Appendices Q and R). 5 panellists involved in the think-aloud protocol (TAP) stage completed a think-aloud evaluation form (see Appendix S)</td>
</tr>
<tr>
<td></td>
<td>Justification for choice of method; detailed account of procedures; feedback from standard-setting panellists</td>
</tr>
<tr>
<td></td>
<td>In this study:</td>
</tr>
<tr>
<td></td>
<td>- Review of potential standard-setting methods outlined in the literature review and a comparative pilot study conducted on two possible methods</td>
</tr>
</tbody>
</table>

There is sufficient and appropriate information for the standard-setting panellists to make a decision

There is sufficient and appropriate information for the standard-setting panellists to make a decision

Standard-setting panellists are confident in the validity of their judgements

Relevant to research question 3 (i.e. How do standard-setting judges view the process and outcome of the standard-setting procedure?)

Feedback from standard-setting panellists

In this study:

- All 18 participants completed a training and final evaluation form (see Appendices Q and R). 5 panellists involved in the think-aloud protocol (TAP) stage completed a think-aloud evaluation form (see Appendix S)
Table 3.2: Argument-based approach and sources of evidence from this study

Firstly, for the assumption “the standards set on the LSP test reflect the language standards in the TLU domain” which is relevant to research question one (i.e. What criteria do standard-setting participants use as a basis for their decisions in judging writing responses and to what extent are these decisions language-based?), the evidence of backing from this study included qualitative commentary elicited from 18 domain expert participants from a range of health professions: general practitioners (GPs), specialists, consultants and medical educators. Secondly, regarding the assumption “standard-setting panellists are oriented to construct-relevant features of the performance” (also relevant to research question one), this study provided qualitative evidence from seven training workshops and five verbal protocols (think-alouds – TAPs). Thirdly, concerning the assumption “variation in qualitative and quantitative judgements of standard-setting panellists is within expectations (or can otherwise be controlled)”, that is related to research question two (i.e. Is there any variability between judges in what they attend to while setting standards?), backing that this study offered is: 1) an analysis of the AJM cut score data and use of subsequent many-facet Rasch program FACETS (Linacre, 2017); and 2) a thematic analysis of the workshop discussion and verbal protocol qualitative data using NVivo software. Next, for the assumption “there is sufficient and appropriate information for the standard-setting panellists to make a decision”, which is relevant to research question three (i.e. How do standard-setting judges view the process and outcome of the standard-setting procedure?), qualitative evidence was gathered from the seven training workshops and five verbal protocols (think-alouds – TAPs) sessions. For the
assumption that “standard-setting panellists are confident in the validity of their judgements” (also relevant to research question three), qualitative evidence was obtained by asking all 18 participants to complete a training and final evaluation form (see Appendices Q and R). Also, the five panellists involved in the TAP stage were requested to complete a think-aloud evaluation form (see Appendix S). Lastly, concerning the assumption “standard-setting procedures are suitable and consistently applied (also relevant to research question three) the following evidence was provided: the AJM method was justified as the chosen method in this study’s method chapter and a comparative pilot study with an alternative method was conducted. A detailed account of the procedure for the trial and main study were given (with procedural refinements) in the previous method sections 3.5 and 3.6. Furthermore, all 18 participants completed a training and final evaluation form (see Appendices Q and R) and five panellists involved in the TAP stage completed a think-aloud evaluation form (see Appendix S).

3.8 Chapter summary

This chapter initially presented the methodological approach of this study. Secondly, two potential alternate methods for use in the main study were briefly described. Thirdly, a report of pilot standard-setting workshops was provided which informed the choice of the final method to be employed. Next, the results from the trial of a qualitative data elicitation procedure (think-aloud protocol (TAP)) was offered. Then, a description of the main study was given including the participants, ethical clearance, instruments and overall procedures. Lastly, an argument-based approach to LSP test validity (as proposed by Knoch and Macqueen, in preparation) and the related evidence that was gathered in this study was illustrated.
Chapter 4: Results

4.1 Introduction

This chapter presents the results of the qualitative and quantitative data analysis. The initial section describes the qualitative findings gathered in the seven workshops and five think-aloud protocol (TAP) sessions. The next section reports participants’ feedback on the standard-setting procedure via their final evaluation comments. The last section states the quantitative results and new cut scores.

4.2 Qualitative results – workshops/TAPs

This section describes the qualitative results from data collected in the seven training workshops and five TAP sessions outlined in the method chapter. The themes that arose from both the workshops and TAPs are presented concurrently. These themes are introduced in parallel as similar aspects were mentioned by participants during both the workshops and TAP sessions. Hence, for ease of comparison, the themes are reported at the same time. This results section develops a preliminary response to three research questions. Firstly, research question one (i.e. What criteria do participants use as a basis for their decisions in judging writing responses and what extent are they language based?). This first research question also corresponds with two assumptions from Knoch and Macqueen’s (in preparation) argument-based validity framework for an LSP test: “the standards set on the LSP test reflect the language standards in the TLU domain” and “standard-setting panellists are oriented to construct-relevant features of the performance”. Furthermore, a response to research question two (i.e. Is there any variability between judges in what they attend to while setting standards?) is also attended to in this results section. Research questions two aligns (in part) with the assumption from Knoch and Macqueen’s (in preparation) argument-based framework “variation in qualitative and quantitative judgements of standard-setting panellists is within expectations (or can otherwise be controlled)”. The final section of the chapter attempts to answer research question
three (i.e. How do standard-setting judges view the process and outcome of the standard-setting procedure?) which aligns with the assumption “there is sufficient and appropriate information for the standard-setting panellists to make a decision” from Knoch and Macqueen’s (in preparation) argument-based framework.

This section provides a comprehensive description of themes that were identified by the thematic analysis. The nine main themes are: ‘Performance Level’ (PL) judgement, ‘Task Fulfilment’ (TF), ‘Content’ (C), ‘Organisation’ (O), ‘Expression’ (E), ‘Presentation’ (PRES), ‘Professionalism’ (PRO), ‘Audience Recognition’ (AR) and ‘Other’ (OTH). Eight of the nine main themes were further sub-divided into sub-themes as seen in table 4.1 below.

<table>
<thead>
<tr>
<th>Theme</th>
<th>Sub-theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘Task Fulfilment’ (TF)</td>
<td>‘Quality’ (Q)</td>
</tr>
<tr>
<td></td>
<td>‘Clarity of Meaning’ (M)</td>
</tr>
<tr>
<td>‘Content’ (C)</td>
<td>‘Purpose’ (PURP)</td>
</tr>
<tr>
<td></td>
<td>‘Case Note Content’ (CNC)</td>
</tr>
<tr>
<td></td>
<td>‘Patient Identification’ (PID)</td>
</tr>
<tr>
<td>‘Organisation’ (O)</td>
<td>‘Discourse Structure’ (DS)</td>
</tr>
<tr>
<td></td>
<td>‘Prioritizing’ (PR)</td>
</tr>
<tr>
<td>‘Expression’ (E)</td>
<td>‘Language’ (L)</td>
</tr>
<tr>
<td></td>
<td>‘Conciseness’ (CON)</td>
</tr>
<tr>
<td></td>
<td>‘Professional Tone’ (PT)</td>
</tr>
<tr>
<td>‘Presentation’ (PRES)</td>
<td>‘Layout’ (LO)</td>
</tr>
<tr>
<td></td>
<td>‘Text Qualities’ (TQ)</td>
</tr>
<tr>
<td>‘Professionalism’ (PRO)</td>
<td>‘Clinical Competency’ (CC)</td>
</tr>
<tr>
<td></td>
<td>‘Patient Awareness’ (PA)</td>
</tr>
<tr>
<td>‘Audience Recognition’ (AR)</td>
<td>‘Audience Awareness’ (AA)</td>
</tr>
<tr>
<td></td>
<td>‘Effort’ (EF)</td>
</tr>
<tr>
<td></td>
<td>‘Handoff’ (HO)</td>
</tr>
<tr>
<td>‘Other’ (OTH)</td>
<td>‘Test Candidate Background’ (TCB)</td>
</tr>
<tr>
<td></td>
<td>OET ‘Task Comments’ (TC)</td>
</tr>
</tbody>
</table>

Table 4.1: Themes and sub-themes

Transcription conventions were previously outlined in methods section 3.6.4.5.1 – qualitative analysis procedure. In addition to this, when citing from the transcripts and in order to retain the anonymity of participants, in this section they are referred to by a unique identifier (i.e. J1 = Judge 1). The researcher is denoted by ‘researcher’.
4.2.1 Performance level judgement

The first theme presented in this results section is ‘Performance Level’ (PL) judgement. This theme refers to standard-setting participants’ summative final decision about a test candidate’s referral letter. This overall judgement corresponds to the seven categories from the AJM method – STRONG to UNSATISFACTORY (as described in the method chapter). These ultimate decisions were usually (but not always) made at the end of discussing/thinking-aloud about a test candidate sample (for both the workshop panellists and think-aloud participants). The following examples illustrate participants giving an overall ‘Performance Level’ (PL) judgement:

This letter is “between Strong and Competent”. It’s all that you need, it’s not, I wouldn’t call it “Strong” or outstanding in anyway, but it’s all that you need [J6, TAP – 277].

I first marked it “Not Yet Competent”, then I read it again and went back and went, no. I put it lower than that [“between Not Yet Competent and Unsatisfactory”] [J11, WK7 – 578].

Standard-setting participants utilized a range of criteria to inform their overall ‘Performance Level’ (PL) judgement of a test candidate’s letter – these are explained in detail in the following sections. In addition, a comprehensive model of how these themes impact on a participants’ final ‘Performance Level’ (PL) judgement are further illuminated in the discussion chapter.

4.2.2 Task fulfilment

The second theme explained in this results section is ‘Task Fulfilment’ (TF). In the process of coding the data, this theme name was chosen to encapsulate the two sub-
themes of overall ‘Quality’ (Q) and ‘Clarity of Meaning’ (M). These two sub-themes, their differences, and examples from the data are shown below.

The first sub-theme encompassed by the theme of ‘Task Fulfilment’ (TF) is participants’ view of the overall ‘Quality’ (Q) of the letter (summative). All standard-setting participants regularly made references to the global quality of test candidate letters, but this was not as specific as a ‘Performance Level’ (PL) judgement. General ‘Quality’ (Q) was commented on by participants when evaluating whether the test candidate had effectively provided adequate information to the recipient of the letter as effectively as possible. These comments were sometimes made by panellists after only having read a few lines or the first section of a letter. Remarks about quality were also often given after participants had concluded reading a letter. Examples of comments related to the overall ‘Quality’ (Q) of a letter follow:

I think this is overall pretty good [J6, TAP – 151].
I think this is quite a good one [J12, TAP – 180].

A second sub-theme of ‘Task Fulfilment’ (TF) is participants’ assessment of the overall ‘Clarity of Meaning’ (M) (i.e. the clarity and intelligibility of a test candidate’s letter being conveyed/or not). If meaning had not been communicated clearly or effectively enough the purpose of the letter would not achieve its task’s aim of transferring information to the receiver. Overall ‘Clarity of Meaning’ (M) was raised by all standard-setting participants. Comments related to meaning were frequently connected and associated with the theme of ‘Expression’ (E) (see section 4.2.5). Panellists often referred to instances of a test candidate’s poor use of expression, but the meaning of the letter was still understandable/intelligible. The converse was also mentioned by participants. Instances of clarity of meaning being affected or not by a test candidate’s use of ‘Expression’ (E) follow:

There’s quite a lot of grammatical and spelling errors so you know on that level, I didn’t want to put them down too much because I understood everything that was there [J8, WK5 – 304].
And then as far as the English expression, when I was thinking about that you just don’t want any ambiguity. You don’t want there to be something written that you’re thinking what did they mean by that? I mean, I think some minor errors of grammar or spelling where it’s still obvious what the meaning is, it’s probably acceptable because I probably do that as well. But I don’t want any of the meaning, like in the person reading the letter to go what do they mean by that? [J15, WK3 – 15].

In summary, ‘Task Fulfilment’ (TF) incorporated the two sub-themes of ‘Quality’ (Q) and ‘Clarity of Meaning’ (M). Participants made global references to these aspects that they value. These features were often linked with further qualities of importance such as ‘Expression’ (E), which are further described below.

4.2.3 Content

The third theme described in this results section is ‘Content’ (C). Through the coding process, this theme designation was selected to capture the three sub-themes of ‘Purpose’ (PURP), ‘Case Note Content’ (CNC) and ‘Identification’ (ID). These three sub-themes, their differences, and examples from the data are given.

The first sub-theme ‘Purpose’ (PURP) refers to the overall purpose of the letter being evident/or not. Standard-setting participants considered this to be an important criterion of a test candidate’s writing response. This theme was often connected with the overall theme of ‘Organisation’ (O) and in particular with the sub-theme of ‘Prioritizing’ (PR). Judges frequently expressed that the purpose of the referral letter should be stated prominently and up front in the opening paragraph. The following are examples of panellists commenting on test candidates’ suitable identification of ‘Purpose’ (PURP) in the letter:

I’m not finished with this one yet, but to start with, the beginning is really good, it’s very clear, the person says clearly what the referral is for [J18, TAP – 11].

I really like the opening statement that says initially what the referring doctor is concerned about and where. The purpose is quite clear [J10, WK6 – 166].

On the other hand, inadequate references to a lack of ‘Purpose’ (PURP) in a test candidate’s referral letter were also made by participants:
It has to say what it’s about in the first sentence. Why the referral’s being made and the purpose of it very clearly up front. I am referring this person because they have a diagnostic or management problem or whatever [J8, WK5 – 55].

It’s not clear to me at this point why this person's referring the patient to the surgeon [J9, TAP – 517].

In addition, a lack of any obviously stated ‘Purpose’ (PURP) in the letter may have an effect on a judge’s overall performance level categorization as seen in this example:

Researcher – ANY OTHER THOUGHTS ABOUT WHAT WOULD PUSH IT INTO THAT “NOT YET COMPETENT” AREA?
J6 – I think the opposite of minimally competent, you know not, not knowing why the referral was sent [J6, WK1 – 110].

The sub-theme of ‘Purpose’ (PURP) was joined with the code ‘Urgency’ (UR) from the initial ‘Data Dictionary’ (see Appendix U). Urgency of the case presented being evident/or not in the letter was mentioned in a limited number of instances by participants. However, from the analysis, these two codes were often mentioned in the same utterance by panellists. Hence, the code ‘Urgency’ (UR) was incorporated under the sub-theme of ‘Purpose’ (PURP) as shown in the following example:

But I think if it would make it difficult for the person receiving the letter to understand what’s actually happening to that patient and how urgent that problem might be, I think if there’s issues with either of those things then for the purpose of the letter, that’s a problem [J15, WK3 – 172].

The second sub-theme included under the theme of ‘Content’ (C) is the more specific ‘Case Note Content’ (CNC). After an analysis of the qualitative data, this sub-theme was combined with several codes from the original data dictionary to also incorporate the following codes shown in table 4.2 below:

| ‘Relevant Content’ (RC)/’Irrelevant Content’ (IC) (redundancy of ideas) from case notes | Possible ‘Treatment Plan’ (TP) evident/or not |
| ‘Content Accuracy’ (CA) and transcription of case note content | ‘Current Medication’ (CM) listed/or not |
| ‘Information Missing’ (IM) from case notes | ‘Regular Medication’ (RM) listed/or not |
| ‘Presenting Complaint’ (PC) evident/or not | Relevant past ‘Medical History’ (MH) evident/or not |
| ‘Clinical Examination’ (CE) evident/or not | Relevant past ‘Social History’ (SH) evident/or not |
‘Diagnosis’ (D)/provisional diagnosis evident/or not

Table 4.2: Combined codes under ‘Case Note Content’ (CNC)

‘Relevant Content’ (RC) from the patient case notes (constructed as a prompt for the OET Writing task rather than actual medical records) being evident/or not was a frequently commented on facet of a letter by standard-setting judges. This could have an effect on their overall performance level categorization as the following example shows:

So yes, it’s “Competent” in that it conveys all the relevant information and it’s not sort of below that standard I think [J9, WK1 – 228].

On the other hand, ‘Irrelevant Content’ (IC) (redundancy of ideas) from the patient case notes could also have an impact on a participant’s view of the letter as seen here:

I guess it’s partly depending on whether they feel they have to include all of this or whether they can leave some of it out, because some of it isn’t... the surgeons won’t care what the pulse rate is, really, or what the height and weight is but they’ve put that in. But, you know, it’s not wrong to put it, but, again, it’s maybe hiding the more relevant facts, just bogging it down with information that’s not so essential [J15, TAP – 273].

In addition, many of the judges mentioned ‘Content Accuracy’ (CA) and transcription of case note content as being important:

**J12** – He said he was ‘42’ which struck me and... [J12, WK1 – 312]

**Researcher** – WHAT WAS THE PROBLEM THERE?

**J9** – He got the age wrong. It’s actually his address number, ‘42’ [J9, WK1 – 316].

Furthermore, ‘Information Missing’ (IM) from the case notes and being excluded in the letter concerned a majority of the panellists which could also affect their overall performance level judgement:

**J12** – So I’d say, “Between Competent and Not Yet Competent” because he should include all the information that’s here.

**Researcher** – INFORMATION FROM THE CASE NOTES?

**J12** – Yes, the case notes [J12, TAP – 354].
Table 4.3 below shows further examples of ‘Case Note Content’ (CNC) being included or not in test candidates’ referral letters. This could have an impact on a standard-setting judge’s overall performance level categorization.

<table>
<thead>
<tr>
<th>Sub-theme</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘Presenting Complaint’ (PC) evident/or not</td>
<td>I think what they’ve actually put about the presenting issues is good [J15, TAP – 411].</td>
</tr>
<tr>
<td>‘Clinical Examination’ (CE) evident/or not</td>
<td>The examination is all done, the description is good [J12, TAP – 413].</td>
</tr>
<tr>
<td>‘Diagnosis’ (D)/provisional diagnosis evident/or not</td>
<td>And, again, right up front, the first paragraph, they’ve said what they suspect is the diagnosis, which I like, just for clarity [J15, TAP – 96].</td>
</tr>
<tr>
<td>Possible ‘Treatment Plan’ (TP) evident/or not</td>
<td>He’s telling the surgeon what he thinks should be done which is reasonable. He explained that he thought the surgeon would most likely remove it [J12, TAP – 246].</td>
</tr>
<tr>
<td>‘Current Medication’ (CM) listed/or not</td>
<td>I probably don’t feel I can quite say “Competent”, because I think that is important, not having the current medications [J15, TAP – 417].</td>
</tr>
<tr>
<td>‘Regular Medication’ (RM) listed/or not</td>
<td>Goes on to mention regular medicines. That’s fine [J12, TAP – 103].</td>
</tr>
<tr>
<td>Relevant past ‘Medical History’ (MH) evident/or not</td>
<td>Ideally if someone’s got significant past history that’s important. The fact that he’s had a previous lesion on that, on another leg, that’s important [J4, WK4 – 45].</td>
</tr>
<tr>
<td>Relevant past ‘Social History’ (SH) evident/or not</td>
<td>A bit of social history so I know who this man is, that’s all good [J9, TAP – 404].</td>
</tr>
</tbody>
</table>

Table 4.3: Further examples of ‘Case Note Content’ (CNC)

A third sub-theme of ‘Content’ (C) was ‘Identification’ (ID). Once more, following the data analysis, this sub-theme was combined to include three codes from the initial data dictionary. The first of these, ‘Patient Identification’ (PID), included a patient being mentioned/or not by name, age or date of birth, address or a combination of
these. A second was ‘Doctor Identification’ (DID) being evident/or not and a third ‘Receiver Address’ (RA) being included/or not.

During the coding process, ‘Patient Identification’ (PID) being mentioned as evident/or not by name, age or date of birth, address or a blend of these was referred to as particularly important by all 18 participants during the workshops and think-aloud sessions. In the OET Writing test criteria currently in use, ‘Patient Identification’ (PID) is assessed as part of the criterion ‘Presentation Features’ (spelling, punctuation, layout). Even though this element is rated as part of ‘Layout’ in the present OET, as this code was regularly referred to by standard-setting participants as an essential constituent of a referral letter, it was included under the theme ‘Content’ (C). Two examples of the lack of ‘Patient Identification’ (PID) being of concern to judges are:

He doesn’t identify the patient. So as one of our colleagues said last time, if this fell on the floor it would be very hard to find out who the patient was. I think that’s a major issue [J12, TAP – 81].

This one, common problem of lacking the date of birth, or address, or other identifying details regarding the patient [J18, TAP – 90].

Furthermore, the inclusion or not of ‘Patient Identification’ (PID) could have an impact on standard-setting panellists’ overall performance level judgements as seen here:

You know the demographic data of the patient, you know just really fundamental stuff, leaving them out, that would be, I would put that as “Unsatisfactory” [J6, WK1 – 160].

For a further analysis of the importance and weight standard-setting judges give to ‘Patient Identification’ (PID) see discussion section 5.3.2.2.

The second sub-theme that was pooled from the initial data dictionary under the theme ‘Identification’ (ID) was ‘Doctor Identification’ (DID) being evident/or not. Again, in the OET Writing test criteria presently being used, doctor identification is rated as an element of the criterion ‘Presentation Features’ (spelling, punctuation, layout). ‘Doctor Identification’ (DID) was mentioned by less than half of the 18 judges overall. However, the following examples show that a lack of ‘Doctor Identification’ (DID) in a test candidate’s letter was an issue for some participants:
Okay, so it is really important that I know who the letter is from and I have no idea who the letter is from. Now I understand this is an exam... [J1, WK2 – 166].

Who is writing the letter, their job title, their address, their contact telephone number and possibly even an email address? [J5, WK5 – 34].

The third sub-theme that was combined from the initial data dictionary under the theme ‘Identification’ (ID) was ‘Receiver Address’ (RA) being included/or not. Once more, in the current OET Writing test criteria, a receiver’s address being present or not is assessed as a part of the criterion ‘Presentation Features’ (spelling, punctuation and layout). The receiver’s address in some form was always included in the letter by test candidates and this code was referred to a limited number of times by two of the 18 judges overall:

This is addressed to the surgeon, surgeon's address [J9, TAP – 177].

They go to great lengths to write in his name and address [referring doctor] which is important for when a referral letter goes back, but the most important thing is the patient’s details which again he leaves out [J12, TAP – 127].

To sum up, it seems that standard-setting judges are influenced by various content features of the writing task when making their overall performance level judgement, including ‘Purpose’ (PURP), ‘Case Note Content’ (CNC) and particularly ‘Identification’ (ID). These are all significant qualities in a writing sample for participants when deciding on an overall performance level categorization.

4.2.4 Organisation

The next theme presented in this results section is ‘Organisation’ (O). In the process of coding the data, this theme name was chosen to capture the two sub-themes of ‘Discourse Structure’ (DS) and ‘Prioritizing’ (PR). As noted in method section 3.6.4.5.1, during the coding validation procedure it was seen that participants made a clear distinction between these two sub-themes. The reason for drawing a division between these two themes is further explained below.

The first, ‘Discourse Structure’ (DS), is related to the general organization and sequence of ideas, stages, tasks, actions and processes of a test candidate’s referral
letter. After an analysis of the qualitative data, this sub-theme was combined with other codes from the data dictionary to also include the following codes: ‘Paragraphing’ (P) of ideas evident/or not (i.e. general comments related to the use or not of paragraphs); use of ‘Logic’ (LOG) evident/or not; and ‘Flow’ (F) of the letter (i.e. general comments related to the overall flow of the letter). Analysis of the transcripts revealed that these codes were used less frequently than the overall code/sub-theme of ‘Discourse Structure’ (DS), and as this theme encapsulated these references, they were collapsed into this one overarching theme.

The general organization and structure of test candidates’ writing responses was a significant criterion for judges when evaluating the quality of the OET samples. The following is an example of a participant commenting on a test candidate using appropriate ‘Discourse Structure’ (DS) in their writing response:

And the structure and the flow of the information so that in reading it you can you know, see in a standard way, you know, that it’s conveyed in the correct sort of order and it’s not all over the place [J9, WK1 – 76].

This next example demonstrates the effect of poor structure and organization on a judge’s view of a test candidate’s writing standard:

It's not quite as well-organized and there's a few places I got caught with reading it. Even with having such a big paragraph as a second paragraph could be probably more succinctly organized into different chunks of information [J15, TAP – 228].

Furthermore, ‘Discourse Structure’ (DS) was important for standard-setting panellists when forming their overall performance level judgements. This could have a positive impact as the following example shows:

Maybe it's more “between Strong and Competent”, because it's fairly clear and well-organized and reads fairly well. Yeah, I think I'll give him a “between Strong and Competent” on that one [J15, TAP – 134].

In this case, the judge has moved their performance level higher from “Competent” to “between Strong and Competent” because of strong ‘Discourse Structure’ (DS). Conversely, a judge may waver with their final judgement due to poor structure. A test candidate’s use of logic and reasoning may also have an effect as in this example:
I'd probably put between “Competent and Not Yet Competent”. It's very close. If we're going to be referring, I think having slightly sort of better language and sort of clunky ordering of things and reasoning even, the sort of communicating I would expect it to be is tidier and clearer [J9, TAP – 283].

As stated, the code ‘Flow’ (F), as described in the original data dictionary was subsequently subsumed under the sub-theme ‘Discourse Structure’ (DS). Some judges also mentioned that the flow of the text could have an effect on their overall performance level judgement as this example highlights:

This is good. This one reads and flows very well. I don’t have any difficulty. I would say this is “between Strong and Competent” [J1, WK1 – 256].

Again, as mentioned, the code ‘Paragraphing’ (P) from the initial data dictionary was pooled with the sub-theme ‘Discourse Structure’ (DS). Once more, the presence or absence of appropriate paragraphing could affect some participants’ performance level categorization as seen here:

I think when I read a letter if there are three different paragraphs one would be an introduction, one’s a background, and one’s the request. That’s almost what I expect and the template I've got in my mind so if something fits immediately visually or on one page then I think that makes it “Strong” because that’s easily communicated [J3, WK4 – 73].

As stated in method section 3.6.4.5.1, judges made a clear distinction between general comments about how the content was structured in the letter, (i.e. ‘Discourse Structure’ (DS)), and the prioritization of the task content (i.e. ‘Prioritizing’ (PR)). An important facet of the letter’s organization for judges is the way in which key details are foregrounded for the reader. Standard-setting participants emphasized frequently that prioritizing crucial details/information (i.e. putting more important information before less important information for the reader) was significant in forming a performance level judgement. The following example illustrates the distinction that judges made between general ‘Discourse Structure’ (DS) and ‘Prioritizing’ (PR). This standard-setting participant has placed particular emphasis on the purpose and key details being placed in a more prominent section of the letter.

It's not too badly organized, but it would be nice for them to be clearer with reason for referral and state their concerns up front [J15, TAP – 76].
Furthermore, another judge again stresses the prominence of the ordering of vital content:

Nothing hidden, that’s important. You don’t want to have to find anything important in the second last sentence, hidden in a way that you miss it. It really has to say everything up front [J8, WK5 – 229].

An effect of poor prioritizing of content can have a direct impact on a judge’s performance level rating of a test candidate as the next example demonstrates:

I wouldn’t say they’re not quite competent yet, I just have to think about how close to competent it is. I think I’m veering to “Not Yet Competent”, just because I feel that having some of the important information up front and grouped together is, sort of shows that medically you know that’s important [J15, TAP – 506].

Another example illustrates that some test candidates who have better prioritization of information may be rated as stronger than those who don’t:

But, again some of the stronger candidates I think up front say, ‘I think they’ve got a recurrent SCC on the leg, that’s why I’m referring them’, whereas here it’s more just descriptive rather than summarizing the issue up front [J15, TAP – 254].

In summary, all standard-setting participants value the general quality of ‘Organisation’ (O) in a writing response when forming overall performance level judgements. As shown, this theme incorporates the sub-themes of Discourse Structure’ (DS) and ‘Prioritizing’ (PR). In addition, as seen in section 4.2.8 below, the organizational qualities of the letter are often linked by participants to the notion of ‘Audience Recognition’ (AR).

4.2.5 Expression

A further theme that is described is ‘Expression’ (E). Through the data coding process, this theme name was selected to encapsulate the three sub-themes of ‘Language’ (L), ‘Conciseness’ (CON) and ‘Professional Tone’ (PT). These three sub-themes, their differences, and examples from the data are presented. As noted in the previous section 4.2.2 – ‘Task Fulfilment’ (TF), test candidates’ use of ‘Expression’ (E) was often linked to the sub-theme of ‘Clarity of Meaning’ (M).
The first sub-theme to be introduced is ‘Language’ (L). After an analysis of the qualitative data, this sub-theme was merged with other codes from the data dictionary. These are described as follows. All participants mentioned accuracy of ‘Grammar’ (G) in the workshops and think-aloud sessions. Examples of judges’ views of test candidates’ use of grammar follow:

So even though the grammar isn’t that good and he gets his capitals mixed up this is quite a good letter [J12, TAP – 136].

I had to use a little bit of effort translating some strange grammar and tone and some words were incorrect [J6, WK1 – 236].

The majority of panellists also commented on test candidates’ use of vocabulary – this was both general and more specific use. In the original coding process, vocabulary was divided into: 1) ‘Language General’ (LG) (i.e. general summative comments about language use); and 2) Vocabulary (V) (i.e. suitable use of vocabulary, appropriateness of word choice and wording being evident/or not). Standard-setting participants often made more general comments related to language use such as:

You know, the English is fine, very respectable [J12, TAP – 251].

This contrasted with more specific remarks about test candidates’ use of suitable vocabulary such as:

They've got this funny kind of use of language. ‘The family history of the patient is irrelevant’, sounds...it's not quite the way we would say it, but it's a pretty minor detail [J18, TAP – 18].

A further code from the initial data dictionary that was merged under the sub-theme ‘Language’ (L) is the suitable use of appropriate ‘Medical Terminology’ (MT) being evident/or not. Again, most participants commented on this aspect as in this example:

Not all the correct language used, like ‘pussy’ you don’t say ‘pussy’, you say ‘purulent’, but that's sort of terminology [J9, TAP – 21].

Another code combined under ‘Language’ (L) was ‘Spelling’ (SP) accuracy. The majority of panellists mentioned spelling accuracy as in this example:

There are a few minor sorts of spelling errors, including the patient’s name, which I just think is only a minor thing, but the patient’s name is pretty important [J7, WK5 – 615].
A final code that was merged under the sub-theme of ‘Language’ (L) was suitable use of ‘Abbreviations’ (AB). Three of the 18 participants commented on their use as in the following example:

I think inevitably in medical letters and communication abbreviations will crop up and I think as a general rule that should be kept at a minimum in letters and communication. Because even though you assume the person receiving it is able to sort of comprehend, these letters really are almost public to the wider medical profession [J3, WK4 – 55].

The second sub-theme encapsulated by the theme ‘Expression’ (E) is ‘Conciseness’ (CON). This included aspects concerning the conciseness, succinctness and efficiency of the letter for the reader being evident/or not. This feature was valued and mentioned by all standard-setting participants. ‘Conciseness’ (CON) was also linked with the themes of ‘Content’ (C) and ‘Organisation’ (O) and in particular the sub-theme of ‘Prioritizing’ (PR). It is also connected to the theme of ‘Audience Recognition’ (AR) as described in section 4.2.8. The following examples illustrate panellists’ comments about a letter’s conciseness:

This is yeah nicely written, it’s succinct, it’s to the point [J12, TAP – 416].

The history’s not bad, it just could be a bit more succinct I would say [J18, TAP – 244].

Furthermore, the quality of ‘Conciseness’ (CON) could affect a participants’ overall performance level judgement as seen in this example:

It's concise, covers what needs to be covered. It's not a distracting document to read. Covers everything I need to know. That's why I'm “between Strong and Competent” [J6, TAP – 178].

A final sub-theme captured by the theme ‘Expression’ (E) is ‘Professional Tone’ (PT) (i.e. the appropriate use of professional tone and register being evident/or not). Around half of judges mentioned this aspect in a test candidate’s referral letter. These examples demonstrate this feature:

And then he said – I don’t like saying: ‘the best assessment management from you’. It’s a little bit sucking up to the person when you don’t need to. I think that tone, I don’t know if it’s a judgement thing. We just need to do our job [J8, WK5 – 599].
And also, I do like the ‘don’t hesitate to contact me if you need further information’. So, it’s opening up the dialogue about communicating that he or she is interested in the patient and wants the outcome to happen. It communicates that they’re serious about it [J10, WK6 – 220].

To sum up, the majority of standard-setting panellists placed some value on the general characteristic of ‘Expression’ (O) in a writing script when establishing overall performance level judgements. As seen, this theme incorporated the sub-themes of ‘Language’ (L), ‘Conciseness’ (CON) and ‘Professional Tone’ (PT). Participants’ comments that related to aspects of ‘Language’ (L) included the codes ‘Grammar’ (G), ‘Language General’ (LG), Vocabulary (V), ‘Medical Terminology’ (MT), ‘Spelling’ (SP) and ‘Abbreviations’ (AB). The sub-theme of ‘Conciseness’ (CON) was valued and mentioned by all standard-setting participants and was also connected with the themes of ‘Content’ (C) and ‘Organisation’ (O) and especially the sub-theme of ‘Prioritizing’ (PR). It was also linked to the theme of ‘Audience Recognition’ (AR). A final aspect of the theme ‘Expression’ (E) was ‘Professional Tone’ (PT) which was mentioned by about 50% of participants as a worthy feature when forming standard-setting judgements.

4.2.6 Presentation

Another theme that is presented in this results section is ‘Presentation’ (PRES). In the process of data coding, this theme name was chosen to capture the two sub-themes of ‘Layout’ (LO) and ‘Text Qualities’ (TQ). The first sub-theme, ‘Layout’ (LO), is related to the overall appropriateness of test candidates’ letter layout/format (i.e. participants’ general comments about the use of standard formal letter layout). This sub-theme was combined with the initial code ‘Presentation’ (PRE) from the original data dictionary. This code is related to the overall appropriateness of letter presentation (i.e. general comments about the overall presentation of the letter such as its messiness/tidiness).

Also, as stated in section 4.2.3 – ‘Content’ (C), during the coding procedure it was seen that participants made a clear distinction between the content theme ‘Patient Identification’ (PID) being mentioned as evident/or not and between the general
layout of a test candidate’s letter. Again, as noted, although patient identification is assessed as a feature of ‘Layout’ in the present OET Writing criteria (and also overlapped during the coding process of this study), as this code was frequently mentioned by standard-setting participants as a vital component of a referral letter, it was incorporated under the theme ‘Content’ (C). For further discussion of the aspect of the crossover between ‘Patient Identification’ (PID) and ‘Layout’ (LO) see discussion section 5.3.2.2.

General ‘Layout’ (LO) aspects were commented on by approximately half of all participants as this example shows:

The first thing I see is it's fairly neat. The layout's quite tidy. It's easy to read. [J6, TAP – 6].

Most panellists stated that general ‘Presentation’ (PRE) issues were of some importance such as in this example:

I think there are some courtesy issues there. For example, you don’t want a piece of scrappy paper, you don’t want dirty paper. It may not be the referrer’s fault but you certainly want something which looks clean and professional [J1, WK2 – 111].

The second sub-theme, ‘Text Qualities’ (TQ) concerns a number of codes that were combined from the initial data dictionary and include: ‘Handwriting’ (HW), ‘Crossed Out’ (CO) information, ‘Punctuation’ (PU) and ‘Text Spacing’ (TS). A majority of participants commented on test candidates’ ‘Handwriting’ (HW) quality/readability as this example demonstrates:

Handwriting’s a bit difficult to read. The handwriting’s quite distracting [J6, TAP – 187].

‘Crossed Out’ (CO) information (i.e. general comments about words, sentences, paragraphs, sections being crossed out) was mentioned by most participants as in this example:

I think it was pretty good though, because I think if you wrote that out without having all this crossing out, kind of like confusing you, I think it would be “between Strong and Competent” [J2, WK3 – 366].
‘Punctuation’ (PU) and other mechanics (i.e. general comments related to punctuation accuracy) was referred to by a limited number of participants as shown in this example:

Commas and things in strange places which makes reading it not so straightforward [J9, TAP – 85].

‘Text Spacing’ (TS) (i.e. the overall appropriateness/or not of text spacing such as writing on every/second line or adding words above lines) was also only commented on by a minority of judges. However, ‘Text Spacing’ (TS) could influence a participant’s overall performance level judgement as seen in this example:

The reason why I’m not making it “Strong” or “between Strong and Competent” is the messiness of the writing, especially the last paragraph. They’ve got little words stuck in on top of sentences. **There’s a bit of extra effort involved in reading this** [J6, TAP – 246].

This aspect of effort (i.e. sentence in bold in the previous example) is also linked to the theme of ‘Audience Recognition’ (AR) and the sub-theme of ‘Effort’ (EF) which are presented in section 4.2.8.

In summary, the two sub-themes of ‘Layout’ (LO) and ‘Text Qualities’ (TQ) were integrated under the overall theme of ‘Presentation’ (PRES). The ‘Layout’ (LO) of a test candidate’s letter was referred to by approximately half of the standard-setting participants. However, as noted, as references to ‘Patient Identification’ (PID) were made by all participants and was seen as an essential element of a referral letter, it was integrated under the theme ‘Content’ (C). Also, as mentioned, patient identification is rated as an aspect of ‘Layout’ in the current OET Writing sub-test criteria. A discussion of ‘Patient Identification’ (PID) being included as a characteristic of ‘Content’ (C) and not ‘Layout’ (LO) is further explored in the discussion chapter. An additional quality mentioned under the theme of Text Qualities’ (TQ) by most participants was ‘Handwriting’ (HW) and ‘Crossed Out’ (CO) information. ‘Punctuation’ (PU) and ‘Text Spacing’ (TS) was referred to by a limited number of panellists.
4.2.7 Professionalism

The theme of ‘Professionalism’ (PRO) is described next. Through the data coding process, this theme designation was selected to encapsulate the two sub-themes of ‘Clinical Competency’ (CC) and ‘Patient Awareness’ (PA). These two sub-themes and examples from the data are presented below.

The first sub-theme, ‘Clinical Competency’ (CC) concerns a test candidate’s clinical competency and skills being evident/or not (i.e. clinical knowledge and skills demonstrated/or not that may have an impact on the referral letter reader’s understanding of the case). Even though standard-setting participants were specifically instructed to discount a test candidate’s ‘Clinical Competency’ (CC) when forming a final performance level judgement, all panellists referred to this feature. The ‘Clinical Competency’ (CC) of test candidates’ skills and knowledge, as evidenced through their interpretation and understanding of the OET task’s case notes, was an aspect that was mentioned by all participants. An instance of this is shown in the following example. The first sentence in bold relates to the sub-theme ‘Clinical Examination’ (CE) of the case. The final sentence in bold refers to the overall ‘Clarity of Meaning’ (M) for the reader based on the test candidate’s ‘Clinical Competency’ (CC) (or lack of):

‘Examination showed no signs of metastases’. I don't really know from the information that you can say that. The information we're given doesn't mention lymph nodes, but I guess they are presuming that because the rest of the exam is normal there's no sign of metastasis. But that's a bit of a medical leap, I think. So that might be misleading for the person reading this [J15, TAP – 105].

Another example illustrates the point of the effect of a test candidate’s demonstration of ‘Expression’ (E) and ‘Content’ (C) and its impact on a participant’s view of their ‘Clinical Competency’ (CC). This ultimately may affect a standard-setting panelist’s final judgement:

I think when you start getting irrelevant pointless information then you start to think is the person medically sound in their assessment more than just a language problem [J3, WK4 – 244].
A final example shows the impact of ‘Clinical Competency’ (CC) (or a lack of) affecting a standard-setting participant’s overall view of a test candidate’s performance level:

It’s bordering on sort of clinical stuff because there’s a lot of irrelevant material that’s in it as well that the surgeon wouldn’t particularly want to know or it doesn’t give a reason. It doesn’t communicate to me why they’ve put it down. They’ve just directly taken ‘the temperature is 36 – 37’. There’s no interpretation that the patient was not ‘febrile’ or there’s no indication of any systemic involvement at this stage. It’s just kind of taking that and putting it in because it might be important, but they don’t really understand how that is [J10, WK6 – 232].

The second sub-theme, ‘Patient Awareness’ (PA) relates to an awareness of the patient being evident/or not (i.e. an awareness of/sensitivity to a patient’s situation or concerns). This aspect was mentioned in a limited number of cases by participants. The following examples demonstrate this feature:

And the bit about saying, ‘well I counselled him because he was anxious’. I think that’s reasonable [J18, TAP – 135].

And it concerned me a little bit that there’s these statements ‘reassuring, reassure him’. I assume that’s past tense, ‘reassured him’ that it was localized. Again those, from my end, I’d be a bit concerned and this is bordering on clinical stuff again about how definite they’ve been with that potentially giving wrong information to the patient [J10, WK6 – 309].

To summarize, the two sub-themes of ‘Clinical Competency’ (CC) and ‘Patient Awareness’ (PA) were included under the overall theme name of ‘Professionalism’ (PRO). Although standard-setting panellists were explicitly told to ignore any evidence of test candidates’ ‘Clinical Competency’ (CC) while deciding on performance level judgements, all participants mentioned this quality. This issue is examined further in discussion section 5.3.2.1. A minority of participants commented on ‘Patient Awareness’ (PA) when making a standard-setting judgement.

4.2.8 Audience recognition

An overarching theme that had a connection with many of the previously described themes is ‘Audience Recognition’ (AR). Through the process of coding the data, this theme name was used to capture the original code designation of ‘Audience
Awareness’ (AA), plus it was merged with additional codes from the data dictionary to also include ‘Effort’ (EF) and ‘Handoff’ (HO). These three sub-themes and examples from the data are described below.

The first sub-theme ‘Audience Awareness’ (AA) being evident/or not is related to a test candidate’s awareness of who the letter is intended for and how it might be read/viewed/interpreted by the reader. This was a theme that was remarked upon by many of the participants and also connected to the themes ‘Content’ (C), ‘Organisation’ (O), ‘Expression’ (E) and ‘Presentation’ (PRES). The following example illustrates the point of intersection of ‘Audience Awareness’ (AA), ‘Content’ (C) and ‘Organisation’ (O):

Thing is the stereotype with surgeons is that they like it to be short and sweet, snappy. Why are you sending them? Tell me now. And if there's other medical stuff going on, they're not always so interested [J15, TAP – 282].

A further example shows the connection between ‘Audience Awareness’ (AA) and ‘Expression’ (E):

It’s very obsequious, I mean you know, ‘I’m very confident he’ll get the very best management with you regarding...’ I mean that’s a lack of understanding of the structure and the intent of professional letter writing [J1, WK2 – 289].

Another example demonstrates the link between ‘Audience Awareness’ (AA) and ‘Presentation’ (PRES):

I think there are other structural issues which need to be, which I’ll say I find very frustrating sometimes when I can’t sometimes even read the name of the doctor or can’t find the name of the doctor on the letter because there’s some scribble, there’s no header on the letter. So, they’re courtesy and form issues, and some of them are legal issues, we do need to know who the doctor is that’s referring the patients so that we can communicate with that doctor and appropriate billing can be communicated with the patient, sometimes you can’t do that [J1, WK2 – 118].

The second sub-theme of ‘Effort’ (EF) refers to the general amount of energy needed for the intended reader to deal with the situation of the letter (i.e. the level of effort required for the reader to comprehend and manage the referral situation efficiently and with minimal exertion). This sub-theme also relates to the theme of ‘Expression’ (E) and in particular ‘Conciseness’ (CON) and ‘Language’ (L). It also has a link to the
theme of ‘Organisation’ (O) and especially the ‘Prioritizing’ (PR) and ‘Discourse Structure’ (DS) of the letter’s ‘Content’ (C). Around half of the participants commented on aspects of ‘Effort’ (EF) as seen in these examples:

I think if a sentence is written in a way that makes the reader read twice or three times then that takes up unnecessary time then that’s “Unsatisfactory” [J3, WK4 – 164].

But I still got the point of his letter and I still understood what he’s trying to say so I feel like it’s still an okay letter it’s just that it takes so much more effort to read it that’s all [J7, WK5 – 530].

The amount of ‘Effort’ (EF) involved in reading a test candidate’s letter for a standard-setting panellist could also affect their overall performance level judgement as demonstrated in this example:

I would say “between Competent and Not Yet Competent”, because I’m just putting myself in the mind of the specialist, so when it comes to a referral letter you don’t really want to have to expend too much mental energy into reading what’s on there, and I’ve had to kind of translate what’s in here and pick out what’s important in the referral and what’s not [J6, TAP – 64].

The final sub-theme under the theme of ‘Audience Recognition’ (AR) is ‘Handoff’ (HO). This refers to a sense the referral letter is justified and not just a cursory handoff to the reader. A limited number of participants commented on this aspect as shown in these examples:

Yeah, when you have like no medical information at all and it’s ‘thank you for referring mister so and so, please manage’. What?! That’s useless. It might give you the provider number but that’s about it. Likewise, if you only put skin lesion: ‘skin lesion query SEC’. Again, that doesn’t really help. Then you have to start from scratch going through all the, trying to get all that information [J4, WK4 – 122].

**Researcher** – OKAY, SO WHAT WOULD MAKE A LETTER “UNSATISFACTORY” THEN?
**J5** – Well, I’ve seen letters like ‘please kindly see’. That’s it. And ‘do as needed’.
**J8** – ‘Do the needful’. It’s a very old saying like if somebody can’t breathe you just say: ‘do the needful’. Like it’s up to you now, I’m giving you the patient, you work it out [J5 & J8, WK5 – 255].
In summary, the overarching quality of ‘Audience Recognition’ (AR) and the sub-theme of ‘Audience Awareness’ (AA) was linked to many of the earlier mentioned themes and this aspect was commented on by approximately half of all panellists. The sub-theme of ‘Effort’ (EF) was also referred to by about 50% of participants. The final sub-theme of ‘Handoff’ (HO) was only referred to by a minority of panellists.

4.2.9 Other

A final category that is outlined in this results section is ‘Other’ (OTH). The codes ‘Test Candidate Background’ (TCB) and OET ‘Task Comments’ (TC) could not be accommodated satisfactorily in any of the previously mentioned themes and hence they were categorized as ‘Other’ (OTH).

The first sub-category, ‘Test Candidate Background’ (TCB), is related to the impact of a test candidate’s background on a standard-setting participant’s overall performance level judgement (i.e. whether the test candidate is local vs foreign, experienced vs novice or English as a first language vs English as a second language (ESL) speaker/writer). Around half of the participants remarked on ‘Test Candidate Background’ (TCB) as these examples show:

So again, he writes his opening paragraph, it flows nicely and makes me suspect he’s maybe a native English writer or speaker [J12, TAP – 262].

J2 – Yeah, I think there’s certain keywords with certain things that we associate that have implicit meaning that maybe they, because it’s a second or third language.

J15 – Like coffee ground and vomit. I mean I use that term and you’re already thinking you know what’s going on, yeah?

J2 – I guess it’s a sub-language, isn’t it? [J2 & J15, WK3 – 155].

The sub-category of OET ‘Task Comments’ (TC) is concerned with general comments about the OET task itself and its features (i.e. comments about the task’s relevance, appropriacy and authenticity compared to standard-setting panellists’ views of real-world practice). Around half of participants commented on this aspect as seen in these examples:
Because I was thinking, I don’t know. Do people still handwrite any letters anymore? They either dictate or type them so that probably wouldn’t happen in a real letter, so I just ignored the messiness [J15, WK3 – 358].

I’m thinking of a situation where you are actually in rural areas and maybe you don’t have the software or you are seeing someone in the emergency as a GP covering acute services and you need to [hand] write and can come across those situations. You still have to work like a GP and do letters cause not everyone who comes there is for an emergency [J5, WK5 – 448].

I mean unlike here, in England, when we do GP letters they are actually written by a secretary. You don’t actually write them yourself. So, the English can actually be changed by autocorrect, stick them in, send back to you so you can approve it and sign it [J5, WK5 – 195].

J2 – It’s [this task’s] like taking our consult notes and writing a letter from it... J15 – ...which is not something any of us really do anymore. We just include those notes, like, with a summary opinion probably [J2 & J15, WK3 – 515].

To sum up, the two sub-categories of ‘Test Candidate Background’ (TCB) and OET ‘Task Comments’ (TC) could not be included adequately in any of the earlier stated themes and were therefore coded as ‘Other’ (OTH). Around half of all participants commented on a ‘Test Candidate’s Background’ (TCB). Once more, approximately 50% of participants made general OET ‘Task Comments’ (TC).

4.2.10 Performance level judgements and decision-making

As mentioned in the literature review, an under-researched area in qualitative standard-setting studies concerns participants’ decision-making processes at the actual time of allocating a performance level judgement. This section attempts to illuminate the thinking processes of panellists by providing examples from the qualitative data of participants’ thoughts while they decided on one performance level category or another. This section also highlights the main themes/qualities that participants valued when allocating a performance level allocation. Particular instances are given of judges who wavered in their decision-making and the themes/qualities of test candidate letters that informed their final judgement. In addition, this results section develops a further response to research question 1 (i.e.
What criteria do participants use as a basis for their decisions in judging writing responses and what extent are they language based?) and the corresponding assumptions “the standards set on the LSP test reflect the language standards in the TLU domain” and “standard-setting panellists are oriented to construct-relevant features of the performance” from Knoch and Macqueen’s (in preparation) argument-based validity framework for an LSP test.

One area that participants especially valued when deciding on a performance level judgement was ‘Content’ (C). The inclusion/or not of relevant and essential content from the OET Writing task case notes was considered of vital importance to many panellists. The following examples demonstrate a participant first judging a test candidate’s letter as higher overall and then after reconsideration, lowering their final judgement due to issues with the content:

**J6** – This one is fine. I would make this “Strong”.

**Researcher** – WHY IS THAT?

**J6** – It’s to the point, it covers everything I need to know. He didn’t mention the glaucoma though. Actually, I will move that down to “between Strong and Competent” because he didn’t mention the glaucoma. Clinically it’s probably not that important, but it is an important part of the medical history [J16, TAP – 130].

I first marked it “Not Yet Competent”, then I read it again and went back and went no and put it lower than that [“between Not Yet Competent and Unsatisfactory”]. I found it confusing and concentrating on irrelevant issues like his eyedrops and so on, and not getting to the crux of the issue quickly. [J11, WK7 – 578].

Another key feature of ‘Content’ (C) is the inclusion or not of ‘Patient Identification’ (PID). In this example, the participant (J12) has stated that the lack of identification has caused them to significantly lower their overall performance level judgement:

If he had of identified the patient I would’ve gone “between Strong and Competent” but he didn’t identify the patient so again I’m going to have to say, “Not Yet Competent”. Look if he had of identified the patient I would’ve marked him up. I feel like I’m nit-picking but, I think it’s, this is clinical medicine. It has to be done properly. [J12, TAP – 226].
A further area of importance for participants when forming a final judgement is ‘Organisation’ (O). Panellists sometimes wavered when making a decision based on this feature as seen in the following examples:

I wouldn’t say they’re “Not Yet Competent” yet. I just have to think about whether how close to “Competent” it is. I think I’m veering to “Not Yet Competent”, just because I feel that having some of the important information up front and grouped together sort of shows that medically you know that’s important [J15, TAP – 506].

I actually said, “between Not Yet Competent and Unsatisfactory” and I thought about “Not Yet Competent”, definitely that, but then I did put it down to almost “Unsatisfactory” because it just seemed to be a real mess in terms of structure [J18, WK7 – 801].

A further point that panellists placed value on was ‘Expression’ (E). Some participants showed that their decision-making was affected by a test candidate’s use of expression as in these examples:

I’d probably put between “Competent and Not Yet Competent”. It’s very close. If we’re going to be referring, I think having slightly sort of better language and sort of clunky ordering of things and reasoning even, the sort of communicating is what I would expect it to be tidier and clearer [J9, TAP – 283].

J12 – Technical terminology I think is pretty important in medicine.
J9 – Yep, and they might say ‘blood in the urine’ but it’s not the correct medical terminology but you know what they’re meaning. So, it might be the difference between maybe “Strong” and...
J12 – …minimally “Competent” [J9 & J12, WK1 – 143].

Throughout the procedure of standard setting, participants were required to assess multiple features of a test candidate’s letter at the same time in order to decide on a final judgement. The following example illustrates the multifarious nature of the standard-setting task and the complex thought processes participants utilized while forming performance level judgements.

Researcher – SO WHY DID YOU GIVE THAT ONE “COMPETENT”? 
J3 – After you got us to think about what criteria makes a “Strong”, “Competent”, “Not Yet Competent”, “Unsatisfactory” letter, some of the sort of criteria that I formulated in my mind is that for a “Strong” letter something that is almost perfect grammatically, information request, outline, format.
So, this doesn’t fit for the reasons that there are a few grammatical errors, spelling errors etc.

And for a “Competent” letter my criteria is that it is clear in what it is asking of me. It has structure and it reads like a letter and it’s not something that I have to read over and over again and ultimately, it’s safe for the patient because they’ve communicated medically accurate information and request what’s necessary so I think this fits the bill.

And with “Not Yet Competent” something that I think is missing information, it doesn’t read well, there’s major grammatical errors that makes it difficult slash impossible to read. They’re starting to fit into that build. And “Unsatisfactory” is just pointless, it doesn’t do anything, so I gave this “Competent”.

J4 – Because you know what the problem is, you know what he’s asking of him and there’s background information. And it’s set out like a letter, but it’s still got problems with wording and grammar and...

J3 – Nothing major I think.

J4 – No nothing major but that stops it from going into the “Strong” category [J3, WK4 – 191].

Furthermore, participants on occasion ‘second-guessed’ themselves while standard-setting and had possible doubts about whether their final judgements were precise enough or not. The following example demonstrates this aspect:

J12 – I think this is “between Strong and Competent”. It’s more than “Competent”.

Researcher – CAN YOU BE MORE SPECIFIC WHY?

J12 – Well he’s got all of the relevant details. He’s identified the patient and he’s got all the relevant details. I’m just wondering why I don’t think it’s “Strong”. I haven’t marked anyone as “Strong”, but it’s, I’d have to go back. It seems a little bit awkward in spots.

Researcher – IN WHAT WAY?

J12 – Line 37, ‘left tibular area’ which ‘has got’ I don’t really like the ‘got’. It’s probably being a bit nit-picky. Ah look this is probably “Strong”, it’s at least “Strong to Competent” but I might be being a bit harsh [J12, TAP – 439].

To sum up, the standard-setting procedure necessitated participants to use complex thought processes while making judgements. The procedure of formulating a final performance level judgement and making global decisions ultimately led to some participants to focus on key aspects of a test candidate’s letter and to use their own indigenous assessment criteria to do so. Some participants wavered when making
these decisions and had doubts about their appropriate level of severity or leniency when assessing writing samples. The discussion chapter further explores this decision-making aspect.

4.2.11 Performance level judgements and variability

This section notes some of the areas in which there was evident variability with domain experts’ performance level judgements. This includes four main themes: ‘Patient Identification’ (PID), ‘Organisation’ (O), ‘Expression’ (E) and ‘Presentation’ (PRES). This section is relevant to research question two (i.e. Is there any variability between judges in what they attend to while setting standards?) and the assumption “variation in qualitative and quantitative judgements of standard-setting panellists is within expectations (or can otherwise be controlled)” from Knoch and Macqueen’s (in preparation) argument-based validity framework for an LSP test.

4.2.11.1 Patient identification

One of the main areas of participant variability was the importance of ‘Patient Identification’ (PID) (or lack of) (i.e. patient’s ‘Name’, ‘Date of Birth (DOB)/Age’ and ‘Address’ (or a combination of these)) in test candidate writing samples. An issue that was common for all participants was the significance of patient identification and its inclusion in the letter. As noted, this feature was valued by all judges, but some more so than others. In addition, some participants placed greater weight than others on the inclusion of patient identification and its significance in their formation of an overall holistic performance level judgement. Also, there was inconsistency with some judges in how the lack of patient ID was applied to their overall performance level categorisation.

In the 200 writing samples used in the study, the majority of test candidates included ‘Patient Identification’ (PID) in some form. Different standard-setting participants valued the inclusion (or not) of ‘Name’, ‘DOB/Age’ and ‘Address’ to varying degrees. Many of the participants stated that they thought it was vital to include a ‘Name’,
‘DOB/Age’ and ‘Address’ in a separate subject line. However, others stated that if an ‘Address’ wasn’t included, the minimum requirement would be a ‘Name’ and ‘DOB/Age’. Yet, some participants also said it would be acceptable (to varying degrees) to identify the patient in the first body paragraph or second body paragraph (more unusual) of the letter using a combination of the above. As mentioned though, the majority specified that a ‘Name’ and ‘DOB/Age’ should be stated somewhere in the letter for the patient to be adequately identified. Only four instances were evident from the 200 writing scripts where test candidates identified a patient by using a name alone. The following table 4.4 provides examples of the various inclusions of ‘Patient Identification’ (PID) in the 200 test candidate scripts.

<table>
<thead>
<tr>
<th>Task 1, T1-BE</th>
<th>Task 2, T2-BJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>n = 100</td>
<td>n = 100</td>
</tr>
<tr>
<td>Name, DOB/Age, Address (stated in subject line)</td>
<td>Name, DOB/Age, Address (stated in subject line)</td>
</tr>
<tr>
<td>Name, DOB/Age (stated in subject line)</td>
<td>Name, Address (Age stated in 1st paragraph)</td>
</tr>
<tr>
<td>Name only stated in subject line/1st paragraph, (Age mentioned in 1st paragraph)</td>
<td>Patient ID mentioned elsewhere = 2nd paragraph</td>
</tr>
<tr>
<td>None = Name only (no DOB/ Age or Address)</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>53</td>
<td>23</td>
</tr>
<tr>
<td>1</td>
<td>18</td>
</tr>
<tr>
<td>39</td>
<td>43</td>
</tr>
<tr>
<td>1</td>
<td>NA</td>
</tr>
<tr>
<td>NA</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 4.4: Inclusions of ‘Patient Identification’ (PID)

As mentioned, some of the standard-setting judges deemed it to be vital to include a combination of ‘Name’, ‘DOB/Age’ and ‘Address’ in test candidates’ letters. As shown in table 4.4, this was witnessed in only 18 instances out of the total 200 letters across the two tasks. The following examples illustrate this point:

**Researcher** – **SO YOU’RE SAYING HERE IN THE SUBJECT LINE HE SHOULD’VE IDENTIFIED THE PATIENT?**  
**J12** – He should’ve at least included a date of birth, preferably a date of birth and an address. Yeah initially I didn’t think it was that important, but when you, again if you see it would be very hard to um... If you found this in someone’s... if it was misfiled and you found this in someone else’s file it would be hard to actually identify this patient and I think that’s a major issue [J12, TAP – 140].
J9 – So for that, I would still say it's not, just under, but close ("between Competent and Not Yet Competent"). It's actually quite a well written letter. If I’m going to stick on the point that the address isn't included and I would expect it to be, then I’d probably have to, again, put it just under “Competent”, even though it’s actually a reasonable letter.

Researcher – EVEN THOUGH IT'S GOT THE DATE OF BIRTH?
J9 – Yeah, I mean, I guess, yeah, it's kind of a, it's a bit of a minor bug really, the whole lack of address thing, but if you're actually wanting to communicate with a patient, it's actually fairly essential information in terms of contacting that patient. Because then this surgeon's going to have to go back to the GP and contact him and get that information. I would expect that sort of information to be included in a letter of this form. But none of them seem to include it. We might have one in there [J9, TAP – 204].

Also, as stated, there was some variability to some of the participants’ preference for a combination of ‘Name’, ‘DOB/Age’ and ‘Address’ to always be included in test candidates’ letters. Some participants deemed the minimum requirement necessary for a patient to be sufficiently identified would be a ‘Name’ and ‘DOB/Age’ (preferably in the subject line). 76 out of 200 of the test candidate letters overall included this combination. The following example highlights the reason for this participant stating that ‘Name’ and ‘DOB/Age’ are enough:

Researcher – WHAT ABOUT THE PATIENT IDENTIFICATION?
J6 – I think it's sufficient. Brian Edwards, and date of birth. That's enough for me to check my records. If there's three Brian Edwards it would be very rare for two of them to share the same date of birth. That’s fine with identification [J6, TAP – 41].

This above example was from the think-aloud transcript of Judge 6. However, in comparison to the previous two examples of J12 and J9, their judgement of the inclusion or not of patient ID was somewhat more stringent (i.e. they required a combination of Name’, ‘DOB/Age’ and ‘Address’ to always be included in test candidates’ letters) whereas for J6 a ‘Name’ and ‘DOB/Age’ was sufficient. Yet, in another situation, J6 is somewhat contradictory to what he said previously about a patient ID’s inclusion or not as the deciding factor in a final performance level judgement – this was often the case for J6 as this think-aloud example shows:

The identifying information. I might not just mention this for the subsequent documents, because automatically if there's no date of birth or address, and
In a further example, it seems evident that J6 was rating writing samples erratically and at times inconsistently as the following post think-aloud discussion illustrates:

**Researcher** – *DO YOU THINK THAT MY LEVEL OF PROMPTING WAS ADEQUATE OR NOT? OCCASIONALLY I ASKED YOU, "COULD YOU BE MORE SPECIFIC, OR IS THERE ANYTHING ELSE?" WAS THAT INTRUSIVE, OR NOT INTRUSIVE?*

**J6** – I don’t think so. It made me think harder about why I gave a rating. It did elicit second-guessing, did I make the right decision, and am I being consistent.

**Researcher** – *THE NUMBER OF RATING RESPONSES THAT I GAVE YOU TODAY, DO YOU THINK YOU WERE COMFORTABLE WITH THAT NUMBER?*

**J6** – They started to blur into each other by the sixth or seventh one. Yeah, by about number seven or eight I was going, ‘I’m not being consistent. I’m just doing this very subjectively now’ [J6, TAP – 292].

Also, there was inconsistency with some judges in how the lack of patient ID was applied to their overall performance level categorization. These examples show that J12 was somewhat inconsistent in his approach, but stressed how important a lack of ID was to him:

> So, because of the lack of identification I’d have to say, “Not Yet Competent” but if he had of had that in, it would’ve been “Competent” [J12, TAP – 153].

> Yep “Competent” except for lack of patient identification. So, I’d have to mark it “between Competent and Not Yet Competent” [J12, TAP – 249].

In the first example, J12 states that because of a lack of sufficient patient ID the test candidate’s final performance level judgement has had to be downgraded from “Competent” to “Not Yet Competent”. However, in the second example, J12 has not used his own self-determined indigenous assessment criteria and has marked this test candidate not as harshly with a final judgement of “between Competent and Not Yet Competent”. This indicates that, at times, judges may be applying their own criteria inconsistently or making errors with the AJM standard-setting scale. These examples were from J12’s think-aloud protocol (TAP) and it would be difficult to determine how many more of these errors/inconsistencies were made with the remaining take-home samples that were judged. This would also apply to the other 17 judges who participated in the study.
The following discussion from training workshop seven further illustrates the point that participants sometimes break their own criteria regarding patient ID when making a final performance level judgement:

**Researcher** – OK. “BETWEEN COMPETENT AND NOT YET COMPETENT”. WHY DID YOU SAY THAT?

**J18** – Once again they left out the address.

**J11** – I marked it “Not Yet Competent”. Again, the address.

**J16** – I’m just revising down because I realized this one didn’t have an address. I’m marking “Competent”. I had something above that but there’s no address.

**J17** – Yeah, I think nobody is going to be “Competent” unless they put the address in so that means I’m in the following scale. But I gave it “Competent”. Yeah breaking my own rule about the address [everyone laughs].

**J14** – I think we have to be bit careful about the address thing because maybe they’re instructed not to worry about that or something. I mean that something that’s easily fixed in my view.

**J18** – Yeah also cause if you use computer software it puts it in there for you automatically. That’s why I marked it down because it didn’t have it but most of the software automatically includes all that stuff.

**J17** – But they should know that.

**J18** – They should know that! [J18, J11, J16, J17 & J14, WK7 – 863].

These examples highlight that some participants view the inclusion of an address in patient ID as vitally important. This was despite the fact that they also knew that in a real-world context this type of information was normally auto-populated by computer software. They still considered it important in a simulated LSP test and marked their performance level judgements accordingly. This was done with some variability (i.e. some judges more harshly than others). In addition, some panellists again broke their own set of indigenous assessment criteria such as J17 in the above discussion. This may have further implications for the validity of the final cut scores.

As shown in table 4.4, some participants judged that it was also acceptable for the inclusion of patient ID elsewhere in the letter (i.e. not only prominently in the subject line). 19 of the 200 test candidate letters included a ‘Name’ and ‘Address’ in the subject line but the patient’s ‘Age’ was stated in first paragraph. The following examples show a participant’s view of the patient’s age being included in the body of the letter:

**Researcher** – YOU GAVE IT “BETWEEN COMPETENT AND NOT YET COMPETENT”. ANY REASONS WHY YOU GAVE IT THAT RATING?
J9 – Let me see. Again, lack of contact, I mean lack of details about the patient. You know there is a name and address but no date of birth. Though they included their age in the first paragraph. I don’t know why I was quite so hard on this one [J9, TAP – 344].

J18 – I would say that's a very, that's a “Strong” letter. They've got some identifying details like that they're 65 and their address. The only thing that could make that better would be if at the start they put the patient’s date of birth and address right up near the top [J18, TAP – 27].

Many of the test candidate letters (82 out of 200) had a ‘Name’ only stated in a subject line and an ‘Age’ (not ‘DOB’) mentioned in the first paragraph. As noted, for many of the judges this was not generally sufficient in their view for the patient to be adequately identified. However, some judges regarded including some of the patient’s details in the body of the letter as enough identification as these examples show:

It’s got the patient's name and then their age in the first paragraph. That’s ok [J15, TAP – 232].

They haven't put the patient's name on a separate line or anything to jump out at you, but it's in the first sentence, with the age. That’s good [J15, TAP – 437].

Only one test candidate letter (T1-BE-12) from the 200 samples placed some form of patient ID elsewhere: in this case patient ID was mentioned in the second body paragraph. However, for some judges this was still sufficient as this example demonstrates:

We don't have a patient's address or date of birth, so “Unsatisfactory” if it's not on the envelope or anywhere else. No, there we go. It's halfway down the letter he's decided to put the address. That's a strange place to put it. This letter is “between Strong and Competent”. I haven't gone all the way to “Strong” because it is a bit long and the address was in the wrong place [J6, TAP – 259].

Once more this example illustrates the variability of some participants’ final performance level judgements.

Only four of the 200 writing samples included a name only and no other form of patient ID (i.e. ‘DOB/Age’ or ‘Address’). As these scripts were from task two they were not commented on during the training workshops or think-aloud sessions and therefore no qualitative data was gathered for these samples. Hence it is unclear to
what extent the total lack of patient ID (apart from a patient’s name) had on a judge’s overall performance level judgement.

4.2.11.1.1 Cambridge Boxhill Language Assessment and patient identification

As noted, participants seem to have strong and at times conflicting views (with others and themselves) on the importance or worthiness of the inclusion (to varying levels) of patient identification when making final performance judgements. The OET Writing sub-test’s construct and criteria used by language-trained assessors also have a particular view. Correspondence from a senior rater at Cambridge Boxhill Language Assessment states that in the sample letters that are produced for assessors and is published for candidates' reference, there is almost always some sort of initial patient ID before the body of the letter. Most often this contains the patient's 'Name' and 'Age' (sometimes ‘Date of Birth’ (DOB)). Also, the supplied patient’s address is sometimes included or the initial diagnosis (e.g., Re: Mrs Betty Windsor, DOB 1/4/1925, Dx: Carpal Tunnel Syndrome).

Feedback from Cambridge Boxhill Language Assessment also reports that the inclusion (or not) of patient ID is not a specific criterion that language-trained raters use to assess writing test responses. In the referral letters that the OET assesses, whilst there is no prescribed standard, candidates are expected to be familiar with this aspect of patient ID as part of the criterion 'Presentation Features' and sub-criterion ‘Layout’. Cambridge Boxhill Language Assessment states that there is a variety of the above combinations, reflecting the variety of the candidates' healthcare settings. They also note that patient ID is almost always included, most commonly ‘Name’ and ‘Age’.

Furthermore, in terms of how this is assessed by language-trained raters, Cambridge Boxhill Language Assessment maintains that patient ID is expected to form part of the letter, just as the date, recipient/receiver’s address, and salutations are appropriate for the genre. The senior assessor at Cambridge Boxhill Language Assessment mentions that: “If too many of these elements are missing, a candidate would certainly be marked down in ‘Presentation Features’, which examines ‘Spelling’, ‘Punctuation’ and ‘Layout’. However, as there are three aspects to this criterion, it
may not have a huge impact on the outcome if the patient ID was incomplete or missing. For instance, it may preclude a 6 (on the OET rating scale from 1-6, low to high), but a 5 would still be possible, if all the other elements were well-controlled” (Cambridge Boxhill Language Assessment, personal communication, March 15, 2017).

The standard-setting participants (as doctors working and practicing in real-world situations) may have placed greater importance on the inclusion of patient ID than the test candidates. In addition, standard-setting judges may have ignored the fact that the test candidate’s writing sample was being used as a proxy to determine their professional readiness and may have regarded the letters as an actual real-world artefact where patient ID is considered extremely important. The following example highlights this point:

**J12** – Could I just ask? [J9] thinks it’s very important to give address, date of birth and contact details. Is that something that’s required in this sort of English test?

**Researcher** – WHAT DO YOU CONSIDER? DO YOU THINK IT’S IMPORTANT OR NOT?

**J6** – For 70 (T1-BE-70), I’m thinking if it was a clinical situation, if this was in someone’s file and it fell on the floor I don’t know which file to stick it back in so yeah it would be a fail of the actual content. If I knew whose file it was in it’s fine [J12 & J6, WK1 – 391].

As demonstrated in the qualitative examples in this section, some participants viewed patient ID as inextricably linked to a test candidate’s professional competence rather than communicative competence which the OET Writing task is designed to assess.

### 4.2.11.2 Organisation

Another area where participants showed variability in their decision-making while setting final performance level judgements was ‘Organisation’ (O). This theme was explained in detail in results section 4.2.4. The two sub-themes of ‘Discourse Structure’ (DS) and ‘Prioritizing’ (PR) especially showed variance among panellists and had varying degrees of importance and impact on their final judgements. The following examples illustrate this point that even though the ‘Content’ (C) of the letter
is judged to be acceptable, issues with ‘Discourse Structure’ (DS) and ‘Prioritizing’ (PR) have led this think-aloud participant (J9) to significantly lower their overall judgement:

I think this letter kind of gets across the information, but in a fairly disordered way. It's competent in a sense that the information's been put down on the page, but in terms of what you'd expect of someone writing a referral letter, even at a fairly junior state, I would expect it to be a bit better in terms of structure and prioritizing of information. So, I would probably stick it, maybe “Not Yet Competent”. It's not “Unsatisfactory”, but it's not what I would expect a letter writer to produce [J9, TAP – 96].

However, another judge (J15), when assessing the same test candidate script, has also found issue with the candidate’s use of ‘Discourse Structure’ (DS) and ‘Prioritizing’ (PR) of information, but has not lowered their final judgement to the same extent:

They haven't expressed what's happening as far as the skin lesion goes. But again, some of the stronger candidates I think up front say I think they've got a recurrent SCC on the leg, that's why I'm referring them, whereas here it's more just descriptive rather than summarising the issue up front, but reads fine. So, I think that one is definitely "Competent". You do have to sort of read through this one a bit more carefully to get the information. Some of them are sort of, the way they're presented, just all the salient features jump out at you. This one, it is a little bit mixed in with things [J15, TAP – 254].

Furthermore, some judges (in this example J15) stated that the prioritizing of ‘Relevant Content’ (RC) and a test candidate’s use of ‘Discourse Structure’ (DS) and ‘Prioritizing’ (PR) of information could also affect a reader’s view of test candidates’ ‘Clinical Competence’ (CC) and hence their overall ‘Performance Level’ (PL). The following example highlights this point:

I wouldn't say they're not quite competent yet. I just have to think about how close to competent it is. I think I'm veering to “Not Yet Competent”, just because I feel that having some of the important information up front and grouped together is sort of shows that medically you know that's important [J15, TAP – 506].

This section shows that participants sometimes vary in their views of the overall ‘Organisation’ (O) of a test candidate’s letter and this may have an impact (also to varying degrees) on their final performance level judgement.
4.2.11.3 Expression

A test candidate’s use of ‘Expression’ (E) was another theme that participants showed variability in terms of importance and overall performance level judgements. Many participants stated that a test candidate’s use of ‘Expression’ (E) was often secondary in their formation of a final judgement as these examples show:

In terms of eloquence and use of language I think that is second rate compared to providing accurate information and making their requests or questions known [J3, WK4 – 30].

So definitely “Competent”. I'll actually put “between Strong and Competent” for that one. Because apart from those two little bits with English, everything else is good. And I don't think they're big problems. They don't cause ambiguity. It's still pretty clear what they're meaning [J15, TAP – 163].

However, some participants valued a test candidate’s use of ‘Expression’ (E) more highly than others and showed variability which could affect their overall performance level judgement as these examples demonstrate:

I mean I, in retrospect maybe, it’s English language, so his English language isn’t that good so perhaps now I’m speaking out loud I’d go down to “between Strong and Competent” [J12, WK1 – 212].

I think if a sentence is written in a way that makes the reader read twice or three times then that takes up unnecessary time then that’s “Unsatisfactory” [J3, WK4 – 164].

Some panellists also contradicted themselves with regards to the value and importance of a test candidate’s use of ‘Expression’ (E). In a previous example, J3 expressed that language use was ‘secondary’ in their formation of a performance level judgement. However, in the following instance they state that the use of ‘Professional Tone’ (PT) is an important quality in a letter.

Yeah, a few pleasantries don’t hurt either. I think it’s all part of professional courtesy. I think that’s important in a way. It’s not as important for sort of patient outcome and patient management, but I think when you’re communicating to another medical professional I think there’s a certain standard in it I suppose [J3, WK4 – 93].

This section illustrates that some judges differ in their judgements of the overall ‘Expression’ (E) of a test candidate’s letter and this may have an impact (also to varying levels) on their final performance judgement.
A final theme in which participants showed some variability in their overall decision-making when forming judgements was ‘Presentation’ (PRES) features. For some judges, this aspect was more important than for others. The following examples show that some participants valued the presentation of the letter and sometimes changed their final judgement due to issues with this aspect:

I originally put “Competent” but because of all the crossing out and the arrows I put it back down to “between Competent and Not Yet Competent” so it was just really distracting [J6, WK1 – 330].

J2 – I think it was pretty good though, because I think if you wrote that out without having all this crossing out kind of like confusing you, I think it would be “between Strong and Competent” because I think it’s all there, it’s pretty clear.

Researcher – SO WHAT WAS YOUR FINAL JUDGEMENT?
J2 – I’m going to say “Competent” [J2, WK3 – 366].

The following examples further show that ‘Presentation’ (PRES) features could have an effect on a panellist’s final judgement. In this instance, two judges (J3 and J4) are discussing the same test candidate’s letter and the impact of ‘Presentation’ (PRES) and ‘Audience Awareness’ (AA) on their overall judgement:

J4 – Well I put “between Competent and Not Yet Competent” and I did think about it. But the way it’s set out with the things crossed out and written over and I think that really is not, it’s not good enough. So, it actually is giving you information but in putting all these little bits in it’s like someone’s just handed in some school work and not done a very good job of it. I would just, you’d be starting again because it’s disrespectful to send a letter like that.

J3 – I gave it “Competent”. I think it’s got all the necessary information. It’s got, it’s demonstrating thought processes and is clear in the request and what the referring person wants and I’ve kind of actually read through what’s been crossed out. He didn’t actually need to cross it out, it’s fine if he left it on. It’s just a grammatical error that demonstrates he probably re-read it and thought about it which is fine and I don’t think the crossing out obscured or took away the information side of things. So, I thought it was “Competent” in the sense that it does what the letter intends to do but it’s not “Strong” [J3 & J4, WK4 – 363].

In this example, the ‘Crossing Out’ (CO) of text from the theme of ‘Presentation’ (PRES) has affected J4’s overall performance level judgement. Also, this is linked to ‘Audience Awareness’ (AA) and a lack of respect (professionalism) even though J4 is aware that
the test format is a hand-written test. However, for J3, ‘Crossing Out’ (CO) and overall ‘Presentation’ (PRES) has not affected their overall final performance level judgement. In J3’s case, the demonstration of thought processes such as ‘Logic’ (LOG) and ‘Purpose’ (PURP) were more important. This section shows that some judges vary in their judgement of the ‘Presentation’ (PRES) features of a test candidate’s letter. This could have an effect (also to varying degrees) on their overall performance level judgement.

Participants draw on the OET Writing task case notes and their own professional/clinical knowledge when making final performance level judgements. However, there was some variability and at times inconsistency with how panellists judged the textual features evident in test candidates’ letters. This involved the inclusion or not (to varying degrees) of ‘Patient Identification’ (PID), ‘Organisation’ (O), ‘Expression’ (E) and ‘Presentation’ (PRES). As shown in the qualitative examples from the think-aloud sessions and training workshops, some judges viewed these features at times dissimilarly. As noted, this was to be somewhat expected as participants were expected to apply their own indigenous assessment criteria. However, some variability and inconsistency remained. This could have possible impacts on the final performance standards and the validity of the cut scores. This is further considered in the discussion chapter.

4.2.13 Summary

This section has described and presented the themes and qualities that were commented on and mentioned as important to standard-setting participants when forming overall performance level judgements. It has developed an initial answer to three research questions. First, research question one (i.e. What criteria do participants use as a basis for their decisions in judging writing responses and what extent are they language based?) which also aligns with two assumptions from Knoch and Macqueen’s (in preparation) argument-based validity framework for an LSP test: “the standards set on the LSP test reflect the language standards in the TLU domain”
and “standard-setting panellists are oriented to construct-relevant features of the performance”. Additionally, research question two (i.e. Is there any variability between judges in what they attend to while setting standards?) is also dealt with in this results section which partly corresponds with the assumption from Knoch and Macqueen’s (in preparation) argument-based framework “variation in qualitative and quantitative judgements of standard-setting panellists is within expectations (or can otherwise be controlled)”. Finally, research question three (i.e. How do standard-setting judges view the process and outcome of the standard-setting procedure?), was further attended to which is related to the assumption “there is sufficient and appropriate information for the standard-setting panellists to make a decision” from Knoch and Macqueen’s (in preparation) argument-based framework.

The discussion chapter further examines construct relevant and irrelevant aspects of particular attributes of test candidates’ referral letters valued by domain expert doctors. A model of participants’ standard-setting decision-making processes is also presented.

4.3 Qualitative results – final evaluations

The following section presents the qualitative results of three final evaluations of the standard-setting procedure by panellists. Domain expert participants were asked to complete three evaluation forms (Training, Final and Think-aloud – see Appendices Q, R and S), depending on their involvement in the study, at a number of stages. This results section provides a further response to research question three: (i.e. How do standard-setting judges view the process and outcome of the standard-setting procedure?) which also corresponds with two assumptions from Knoch and Macqueen’s (in preparation) argument-based validity framework for an LSP test: “standard-setting panellists are confident in the validity of their judgements” and “standard-setting procedures are suitable and consistently applied”.

192
4.3.1 Training workshop evaluation

Panellists were asked to complete a ‘Training Evaluation Form’ (see Appendix Q). This was undertaken to ascertain whether judges understood what was required of them to set standards using the Analytic Judgement method (AJM) and to determine whether any further training would be necessary. Participants were asked whether they agreed or disagreed with a number of statements on a 5-point ‘Likert’ scale: ‘strongly agree’, ‘somewhat agree’, ‘neutral’, ‘somewhat disagree’ and ‘strongly disagree’.

All participants agreed that they understood the standard-setting task and that the training workshop sessions were adequate. The majority of panellists (93%) indicated that they clearly understood the performance level categories and their differences including the ‘between’ categories while one participant (J6) was ‘neutral’. All participants confirmed that they understood how to make judgements using the AJM and that they understood how to use the form provided (see Appendix D).

Panellists were also asked about the influence of others on their decision-making when considering performance level judgements during the training workshops. This was not relevant for two judges, J1 and J10, as they could only be present at a training session separately with the researcher due to time and logistical limits. Participants were asked whether the feedback and discussion helped to refine their understanding of the standard-setting process and all agreed. In regard to whether the feedback and discussion helped to refine panellists’ judgements and understanding of the performance levels, the majority agreed (93%) and one panelist was neutral (7%). Participants were also asked if the feedback and discussion would influence their further writing response judgements. Again, most agreed (93%) and one was neutral (7%).
4.3.2 Final evaluation of the standard-setting procedure

Panellists were asked to evaluate the standard-setting procedure overall by completing a ‘Final Evaluation Form’ (see Appendix R). Participants were asked to indicate how influential several factors were when making performance level judgements. This was on a 3-point ‘Likert’ scale: ‘very influential’, ‘somewhat influential’ and ‘not at all influential’. Some of the statements were not applicable for two panellists (J1 and J10), as due to time and logistical restraints they could only be present at a training workshop separately with the researcher.

All participants (100%) said that ‘my experience with the type of writing task from my working life’ and their ‘sense of what test candidates should be able to do in order to be proficient/safe in the workplace’ was influential on their performance level judgements. In terms of ‘participants’ own definitions of performance levels’ all stated it was influential. Regarding ‘other participants’ definitions of performance levels’, the majority (87%) said they were influential, but two participants (J4 and J6) (13%) said that others’ definitions were not at all influential. In relation to how influential the ‘feedback/comments/discussion with other participants’ was, all indicated it to be influential. Also, the ‘writing response judgements/ratings of other participants’ was considered and the majority (75%) stated it was influential, but 25% indicated it was not at all influential.

Panellists were asked to state how useful a number of aspects were when they were judging writing responses. This was on a 3-point ‘Likert’ scale of ‘very useful’, ‘somewhat useful’ and ‘not at all useful’. This was not applicable for two panellists (J1 and J10) for some of the statements as they attended the training individually. 100% of participants stated that ‘practicing the procedure’ was ‘very useful’. Concerning the ‘feedback/comments/discussion with other participants’, all said it was useful. Most participants (87%) found the ‘writing response judgements of other participants’ to be useful and 13% stated it to be ‘not at all useful’.

Finally, all participants were asked to consider the standard-setting procedure overall and whether it was a valuable exercise or not. Panellists indicated the degree to which
they agreed with a number of statements on a 5-point ‘Likert’ scale: ‘strongly agree’, ‘somewhat agree’, ‘neutral’, ‘somewhat disagree’, and ‘strongly disagree’. Most judges stated that they agreed with the statement: ‘I believe that from this study, I have a greater understanding of the importance and value of standard setting’, one participant was ‘neutral’ and one (J14) said ‘somewhat disagree’. In regard to the statement whether ‘the standard-setting procedure was a worthwhile exercise for me’ 100% agreed. Lastly, concerning the statement: ‘I believe that in future, standard-setting sessions should be conducted at regular intervals’, most panellists agreed and two were ‘neutral’.

All standard-setting participants were asked if they had any additional comments or feedback about the standard-setting process. These comments were analyzed and ordered thematically. An initial theme that many of the participants mentioned was the ‘value of standard setting’. J10 pointed out that standard setting was worthwhile: “I came to a strong understanding of the importance and value of standard setting. Very interesting process and really important to standard set to ensure candidates are all marked consistently”.

J15 also indicated that standard setting was important in terms of equality for test candidates and remarked: “I think it’s very important for fairness for candidates”. Furthermore, J5 regarded standard setting as vital and stated: “I believe the standard of English and the understanding of information given (written or otherwise) should be high in the medical profession because errors can be fatal and lead to mismanagement of patients as a whole”.

A necessary component of standard-setting procedures is for participants to use their own ‘indigenous assessment criteria’ when forming performance level judgements. This is based on panellists’ individual knowledge and experience from the workplace. However, some of the participants expressed that under the theme of ‘criteria’, they would have preferred to have been given pre-determined benchmarks or guidance in order to make judgements. The following examples illustrate this point:

J15 – I think that some sort of marking criteria would help as markers may each value different aspects and mark accordingly.
J11 – It would’ve been useful to know the basis of the overall assessment of candidates and what constitutes a ‘pass’ in the assessment.

J8 – I think this exercise would have worked better if there had been optional other presence of errors across the cases for example: i) patient details; ii) synthesis; iii) grammar, spelling; iv) sequencing; v) prioritization.

J17 – The training session was very useful, but perhaps some guidance whether there were/should be fatal errors would make the marking more uniform?

J1 – I believe that certain essential elements of the correspondence should be identified. If these are present then the candidate should be marked up, conversely marked down if absent.

A related theme to criteria was ‘patient identification’ as being evident or not. Many of the participants stated that further guidance would have been useful to indicate how important the presence and especially lack of patient ID had on their performance level judgements. J8 pointed out that:

I feel a major flaw was the issue of patient address (or lack of). In the majority of cases this was omitted and thus the letter had to be marked down. It suggests the candidates worked on the assumption that this information would be provided (e.g., by addition of a sticker). This skewed the results.

J6 also noted that:

It would be worthwhile specifying how important it is for candidates in the test to adhere to business writing conventions (format, demographic details, date) in addition to the letter content.

One of the participants stated that the ‘background of judges’ is an important consideration for a standard-setting panel. J16 stated:

The background of the assessors is a significant factor in standard setting. For example, a GP working in exclusive clinical practice may have a different standard than an academic physician working in a teaching hospital. During the group exercises I was the harshest assessor of these responses, but the following day I was involved in standardization for the Royal College of Physician examinations, where I was the most generous assessor.

Another panelist also noted that ‘further training’ beyond the initial training workshop would have been useful. J2 stated that:
The standard-setting training may have benefited from a further few collaborative marking cases. I felt that at times I judged the proficiency of the sample on content (e.g., if they missed out particularly important information, seemed to infer things that weren’t actually there or weren’t very clear on what they were asking the GP to follow-up) and I know that it was supposed to be more writing style rather than content.

In summary, many of the participants stated that they considered standard setting to be a valuable and important procedure to undertake. However, some felt that distinct criteria would have been useful, rather than having judges rely on their own indigenous assessment criteria while forming judgements. A particular example of this for some judges was the weight given to the presence or absence of patient identification. One of the participants noted the importance of a judge’s background on how they made judgements. Another panelist also expressed the desire for further training with more practice writing responses to compare judgements with other participants.

4.3.3 TAP session evaluation

Five of the 18 judges took part in think-aloud (TAP) sessions with the researcher and thought out loud about the same 12 writing samples. These five TAP participants also filled in a ‘Think-aloud Final Evaluation Form’ (see Appendix S). The think-aloud judges were asked to specify their level of agreement on a number of statements concerning the think-aloud procedure. Participants were asked to state how much they agreed with a number of statements on a 5-point ‘Likert’ scale: ‘strongly agree’, ‘somewhat agree’, ‘neutral’, ‘somewhat disagree’ and ‘strongly disagree’.

All five think-aloud judges indicated that they ‘strongly agree’ ‘the instructions were clear’ and ‘I understood what I was required to do when ‘thinking aloud’’, ‘the level of prompting from the researcher was adequate’, and ‘I had adequate time to speak comprehensively about each writing response’. Four of the five judges said they ‘strongly agree’ that ‘I was comfortable with the number of writing responses I judged and spoke about’ while one judge (J6) indicated ‘somewhat disagree’ with this statement. Finally, the think-aloud participants were asked to consider the statement:
‘the process of ‘thinking aloud’ didn’t alter the way in which I judged the writing responses’. Three (J9, J12, J18) of the five said they ‘strongly agree’, J6 indicated ‘somewhat agree’ and J15 said ‘somewhat disagree’. The think-aloud judges were asked if they had any additional comments about the experience of ‘thinking aloud’ overall, but none were expressed.

4.3.4 Summary

To sum up, as part of a final evaluation of the standard-setting process, participants were asked to complete the forms (Training, Final and Think-aloud – see Appendices Q, R and S), depending on their involvement in the study, at a number of stages. These qualitative results offered an initial response to research question three: (i.e. How do standard-setting judges view the process and outcome of the standard-setting procedure?). This research question also aligns with two assumptions from Knoch and Macqueen’s (in preparation) argument-based validity framework for an LSP test: “standard-setting panellists are confident in the validity of their judgements” and “standard-setting procedures are suitable and consistently applied”. The results of these evaluations are further examined in the discussion chapter.

4.4 Quantitative results

This section introduces the results of the analysis of quantitative data derived from domain experts’ judgements of samples of OET Writing sub-test performance. This quantitative analysis was carried out as part of the examination of research question two: Is there any variability between judges in what they attend to while setting standards? Furthermore, research question two also corresponds (in part) to the assumption from Knoch and Macqueen’s (in preparation) argument-based framework “variation in qualitative and quantitative judgements of standard-setting panellists is within expectations (or can otherwise be controlled)”. In addition, these quantitative results offer a response to research question four: What occupational specific standards (cut scores) do health professionals set on the Occupational English Test (OET) Writing sub-test?
4.4.1 FACETS analysis

As noted in the methods chapter, the data were analysed using the many-facet Rasch program FACETS version 3.8.0 (Linacre, 2017). The Rasch analysis displays the individual facets of enquiry onto a logit scale. This logit scale is a scale on which the various facets in the analysis are positioned and shows how all these facets of the assessment situation relate to each other. In the case of this study, the facets of interest are ‘test examinees’, ‘raters’, ‘test task’ and the ‘AJM rating scale’. The ‘variable map’ or ‘Wright map’ (see figure 4.1) gives a visual overview of the FACETS analysis (Bond & Fox, 2015; Eckes, 2011; McNamara, 1996).

Figure 4.1: ‘Variable’ or ‘Wright’ map

In figure 4.1, in the first column (‘Measure’ or ‘Measr’) displays the common logit scale on which all facets from the analysis (candidates, raters and task) are situated. This acts as a ‘ruler’ and allows areas of enquiry to be located on this scale (McNamara, 1996). The second column (‘Candidate’), shows estimates of candidate ability. In this case, for display purposes, each star (*) represents two test candidates and each dot (.) one test candidate. The test candidate ability measure is ordered from highest to
lowest with the candidate with the highest proficiency being shown at the top of the column. The third column (‘Rater’), displays the comparative severity or leniency of raters. The most severe rater appears at the top of the column while the least severe is at the bottom. The fourth column allows for comparison of the relative difficulty of the two tasks, but for the purposes of the current study, these were considered to be of equal difficulty and therefore both anchored at zero. As noted in the methods chapter, in order for OET Writing sub-test tasks to be fair for all test candidates in each administration, in their construction and development they are designed and intended to be of equivalent difficulty. The final column (‘Scale’), represents the seven AJM performance levels from STRONG to UNSATISFACTORY which, for this analysis, were converted to a 7-point scale from 7 – 1 with seven corresponding to STRONG and one to UNSATISFACTORY.

As seen in figure 4.1 (column 2), the FACETS analysis reveals significant variability in test candidate ability (i.e. the stars (*) and dots (.) representing test candidates are distributed across various points on the vertical ruler from -3 to +4 on the logit scale). This is to be expected as the OET Writing sub-test samples were selected to represent a full range of candidate abilities from STRONG to UNSATISFACTORY (i.e. OET band A – D). Also, as shown in figure 4.1, some raters displayed variability with regards to severity/leniency of their judgements of test candidate responses. One judge in particular (J6) in column three was much harsher than the other raters in how they judged writing competence as can be seen by their positioning well above the other raters on the vertical scale. For a full discussion of this aspect, see next section 4.4.1.1, rater severity/leniency.

4.4.1.1 Rater severity/leniency

An analysis of the standard-setting judges’ comparative rater severity/leniency was conducted. This was performed as part of the investigation of research question two: ‘Is there any variability between judges in what they attend to while setting standards?’ Measures below -1.0 represent lenient raters and measures above 1.0 characterize severe raters (i.e. raters below -1.0 are lenient and those above 1.0 are
severe) (Eckes, 2011). In table 4.5 below, selected columns from the rater measurement report are presented.

<table>
<thead>
<tr>
<th>Judge</th>
<th>Measure logit</th>
<th>Model standard error</th>
<th>Infit MnSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>1.79</td>
<td>0.16</td>
<td>2.07</td>
</tr>
<tr>
<td>11</td>
<td>0.73</td>
<td>0.18</td>
<td>1.10</td>
</tr>
<tr>
<td>18</td>
<td>0.64</td>
<td>0.14</td>
<td>0.85</td>
</tr>
<tr>
<td>16</td>
<td>0.60</td>
<td>0.17</td>
<td>0.90</td>
</tr>
<tr>
<td>14</td>
<td>0.59</td>
<td>0.17</td>
<td>0.93</td>
</tr>
<tr>
<td>10</td>
<td>0.56</td>
<td>0.18</td>
<td>1.06</td>
</tr>
<tr>
<td>5</td>
<td>0.03</td>
<td>0.17</td>
<td>0.65</td>
</tr>
<tr>
<td>8</td>
<td>-0.03</td>
<td>0.17</td>
<td>0.92</td>
</tr>
<tr>
<td>12</td>
<td>-0.05</td>
<td>0.15</td>
<td>0.83</td>
</tr>
<tr>
<td>7</td>
<td>-0.10</td>
<td>0.17</td>
<td>0.67</td>
</tr>
<tr>
<td>4</td>
<td>-0.27</td>
<td>0.17</td>
<td>1.47</td>
</tr>
<tr>
<td>13</td>
<td>-0.28</td>
<td>0.17</td>
<td>0.66</td>
</tr>
<tr>
<td>9</td>
<td>-0.30</td>
<td>0.15</td>
<td>0.72</td>
</tr>
<tr>
<td>17</td>
<td>-0.36</td>
<td>0.17</td>
<td>1.31</td>
</tr>
<tr>
<td>15</td>
<td>-0.75</td>
<td>0.15</td>
<td>0.60</td>
</tr>
<tr>
<td>3</td>
<td>-0.89</td>
<td>0.18</td>
<td>1.17</td>
</tr>
<tr>
<td>1</td>
<td>-0.89</td>
<td>0.18</td>
<td>0.56</td>
</tr>
<tr>
<td>2</td>
<td>-1.03</td>
<td>0.18</td>
<td>0.88</td>
</tr>
<tr>
<td>Mean</td>
<td>0.00</td>
<td>0.17</td>
<td>0.96</td>
</tr>
<tr>
<td>SD</td>
<td>0.70</td>
<td>0.01</td>
<td>0.36</td>
</tr>
</tbody>
</table>

RMSE: 0.17       Adj. S.D: 0.68       Separation 4.07
Reliability (not inter-rater): 0.94       Significance (probability): 0.00

Table 4.5: Rater measurement report

In the first column, judges are shown in descending order from most to least severe. As seen in the second column, the analysis indicated that one of the raters (J6) was considerably more severe than the other standard-setting participants with a logit score of 1.79. This was considerably more severe than the next harshest judge (J11) at 0.73. Conversely, one judge (J2), was slightly more lenient than the others with a logit score of -1.03. This was somewhat more lenient than the nearest other judges with logit scores of -0.89 (J1 and J3). The third column displays the ‘model standard error’. This shows that as the estimates of raters’ severity and leniency are extrapolations from the data set, they may be subject to error (McNamara, 1996).
Linacre (cited in Eckes, 2011) states “the smaller the standard error, the higher the estimate’s precision” (p.54) or that higher precision is more indicative of a rater’s ‘true’ level of severity. For instance, the severity measure of J6 was 1.79 logits, with SE = 0.16. According to Eckes (2011), assuming the distribution is about average, J6’s ‘true’ measure would be anticipated to lie within ±2SE around that estimate 95% of the time, giving an interval of an upper range of 2.11 (i.e. 1.79 + 2 x 0.16) and a lower range of 1.47 (i.e. 1.79 – 2 x 0.16). Eckes (2011) also notes, that all things considered equal, the larger the amount of observations an estimate is founded on, the smaller its standard error will be, as will the associated interval. The fourth column deals with raters’ consistency or ‘fit’ with the model. This gives an indication of the extent to which each element acts in a way that is expected by the model (Bond & Fox, 2015). Rater fit is examined further in section 4.4.1.2.

In addition to examining rater severity/leniency, the FACETS analysis also offers two additional gauges – the separation index and the reliability figure index to indicate the extent of variances between a facet’s features (i.e. the level of difference between features of a facet) (Bond & Fox, 2015). The separation index is the ratio of the adjusted standard deviation (Adj. SD: 0.68) of element measures (in this instance, raters) to the root mean-square standard error (RMSE: 0.17). If the judges were equally severe/lenient, the standard deviation of the judges’ difficulty estimations would correspond to or be less than the mean estimation error of the whole dataset (Bond & Fox, 2015). The rater separation index, however, in this study’s data is 4.07. This shows that the variance between raters is approximately four times the error of estimates. This means that the standard-setting judges are not rating in a similarly severe/lenient way, or in other words, that there is a level of difference and variability amongst them.

The reliability figure offered by FACETS shows the degree to which the analysis consistently differentiates between various stages of difficulty, or severity/leniency, amongst the elements of the facet; in this instance, different raters. According to McNamara (1996), a low reliability score would show the various judges as being equally severe/lenient. However, on this occasion, the reliability figure is 0.94. This
shows that the analysis is quite reliably dividing judges into different levels of severity/leniency. This again means that the standard-setting judges are not judging writing samples in an equally severe/lenient manner and that there is significant variability between them. In summary, the separation index and the reliability measures indicate that the standard-setting judges have not been rating writing scripts in an equally severe/lenient fashion.

4.4.1.2 Rater consistency

Participant consistency can also be measured by the FACETS analysis. This is represented by statistical indications of the degree to which judges used the AJM rating scale in a consistent way. This consistency measure shows the level of ‘match’ or ‘fit’ between what was anticipated by the model and the actual data (McNamara, 1996). In this case, rater fit relates to the degree that a judge is connected to unexpected ratings, summarized over test candidates and the AJM rating scale. Judges who display very high mean square values (infit) are thought to be misfitting (i.e. their ratings do not sit within the anticipated range that the FACETS program expects based on the data provided to it). This typically indicates judges who may be rating inconsistently and could mean, for instance, that a judge may be applying his/her own indigenous assessment criteria to test candidates’ writing samples in different ways or using them haphazardly. This is to be somewhat expected as participants were not given any pre-specified criteria. McNamara (1996) argues that when judges exhibit rating patterns that are ‘erratic’, this makes it problematic to model their characteristics and hence compensate for them. On the other hand, judges with very low infit statistics may be rating more predictably than the program anticipates and therefore show ‘overfit’. This could indicate judges who are overusing particular performance levels and therefore not showing the sort of estimated variation amongst test takers that might be expected (McNamara, 1996). Overfit is less of a problem than misfit as misfitting raters may alter the applicability of the subsequent measures and consequently endanger the validity of the measurement system (Eckes, 2011; Myford & Wolfe, 2003; Wright & Linacre, 1994).
A suitable range of fit (upper and lower limits) needs to be decided on for practical purposes in determining rater consistency. Even though there are no fixed parameters for determining the degree of fit, as a general principle, Linacre (cited in Eckes, 2011) has suggested on a number of occasions an upper limit of 1.50 and a lower limit of 0.50 as “productive for measurement” or as suggestive of “useful fit” (p.58). Lunz and Stahl (1990) also established the lower and upper bounds of 0.5 and 1.5 to be acceptable for mean squares to be used for practical purposes. Others, such as McNamara (1996) and Knoch and McNamara (2015) have specified a narrower range of 1.3 (high) to 0.75/0.8 (low) as ‘usual’. However, Eckes (2011) argues that in practice, the degree that is defined by the upper and lower limits will vary according to the reality of the assessment purpose. For this study, the broader range of 0.5 to 1.5 put forward by Lunz and Stahl (1990) and Linacre (2017) was utilized.

It can be observed in the fourth column in table 4.5 that one of the judges (J6) displayed significant misfit (Infit MnSq: 2.07). This is well above the accepted upper range of 1.5 of Lunz and Stahl (1990) and Linacre (2017). Therefore, J6 may have been allocating performance level judgements inconsistently. Conversely, none of the 18 judges could be determined as overfitting (i.e. not using an acceptable variety of performance level categories).

In summary, in terms of rater severity/leniency, one of the judges (J6) was particularly severe and one (J2) was slightly more lenient than the other participants. The rater separation index figure of 4.07 indicates that the difference between judges is approximately four times the error of estimates. This shows that the participants are not judging writing performances in a correspondingly severe/lenient manner (i.e. there is considerable difference between them). Also, the reliability figure of 0.94 reveals the FACETS analysis is quite reliably separating raters into different levels of severity/leniency (a low reliability figure indicates raters as equally severe/lenient). In terms of consistency, one of the judges (J6) seemed to be making performance level judgements unpredictably. However, the majority of judges appeared to be allocating performance level judgements quite consistently and across a satisfactory range of performance level categories. Rater behaviour such as severity/leniency and
consistency can have implications for the new cut scores and the procedures for dealing with these issues are described below.

4.4.2 New cut scores

As described, the Analytic Judgement method (AJM) was used as the standard-setting method for the study. In this method, writing samples were selected to represent a full range of performance levels and were presented randomly to participants in no specific ordering. As stated, the OET test candidate fair scores (derived from a FACETS analysis of English-language trained professionals’ raw scores) for the study were supplied by Cambridge Boxhill Language Assessment and were not given to panellists (although they were later used for the calculation of cut scores as described below). Instead, participants made their judgements independently of any previous proficiency rating using a designated form, which required them to place the identifier of each response into an appropriate cell from either of the four main categories STRONG, COMPETENT, NOT YET COMPETENT and UNSATISFACTORY or into a ‘between’ category. Cut scores were established by using the fair scores previously assigned by OET raters to each of the scripts that the domain expert judges had allocated to one or other of the between categories. The sum of all the pooled fair scores for each script placed in a between category was averaged to give a mean score. The mean of the fair scores for the performances in each of the between categories became the new cut scores. The cut scores drawn from the OET fair scores were expressed on a scale from 1 to 6.

Before the final cut scores were determined, the above-mentioned factors concerning rater severity/leniency and consistency were taken into account. As seen in the FACETS analysis (see previous section), one of the judges (J6) was noticeably more severe than the other raters with a logit score of 1.79. Additionally, one judge (J2) was somewhat more lenient than the other participants with a logit score of -1.03. As noted, the purpose of the AJM is not for 100% exact rater agreement and this is not to be expected for raters in general and in particular for those making judgements without specified criteria to guide them. However, as Linacre (2017) states, some
raters may need to be eliminated from the analysis (i.e. those who display idiosyncratic profiles which could have a misleading impact on the overall results). The two ‘outlier’ panellists J2 and J6 were hence removed from the final AJM cut score calculation.

In addition, as shown in the FACETS analysis (see previous section), one of the judges (J6) was rating inconsistently or overusing performance levels and was ‘misfitting’. The fact that this judge’s Infit MnSq was above 1.5 (i.e. above the conventional upper range of 1.5 of Lunz and Stahl (1990) and Linacre (2017)) suggested that, as well as being overly severe, as noted above, J6 was allocating performance level judgements unpredictably. This confirmed the decision to exclude rater J6 from the final cut score calculation. As mentioned, a misfitting rater could change the consequences of the ensuing measures and subsequently compromise the validity of the measurement system (Eckes, 2011; Myford & Wolfe, 2003; Wright & Linacre, 1994).

As seen in table 4.6 below, there were differences between the current OET cut scores (OET, 2009) and the new cut scores as determined by medical professionals via the standard-setting procedure described above. The new cut score of 5.26 for ‘between STRONG and COMPETENT’ (the minimum score required for an OET A grade) was slightly less stringent than the current OET score of 5.6. For ‘between COMPETENT and NOT YET COMPETENT’ (the minimum needed for an OET B grade) the new cut score of 5.04 was somewhat higher (and therefore more stringent) than the present OET one of 4.8. The new cut score of 4.77 for ‘between NOT YET COMPETENT and UNSATISFACTORY’ (or the minimum requirement for an OET C grade) was significantly higher (i.e. harder to attain) than the current OET cut score of 4.2.
Table 4.6: New vs current OET cut scores

Another difference between the new and old OET cut scores worthy of note is the range or spread between levels (i.e. the distance of cut scores from each other). The range between the current OET cut scores and the new cut scores is now significantly different. For the current OET cut scores, the gap between OET band A/B and B/C is 0.8 and between B/C and C/D it is 0.6. However, for the new cut scores the difference is far less: between OET band A/B and B/C the gap is 0.23 and between B/C and C/D it is 0.27. Overall, this has major consequences for the candidates represented in the current data set. If the new cut scores as determined by doctors were employed, many of these test candidates would have received a different final grade than the one previously assigned. More precisely, 127 test candidates out of the sample of 200 (64%) would have ended up with a different grade on the OET than the one they had previously received based on the existing cut scores.

<table>
<thead>
<tr>
<th>AJM Category or OET Bands</th>
<th>New Cut Score</th>
<th>Current OET Cut Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Strong and Competent or (A)/(B)</td>
<td>5.26</td>
<td>5.6</td>
</tr>
<tr>
<td>Between Competent and Not Yet Competent or (B)/(C)</td>
<td>5.04</td>
<td>4.8</td>
</tr>
<tr>
<td>Between Not Yet Competent and Unsatisfactory or (C)/(D)</td>
<td>4.77</td>
<td>4.2</td>
</tr>
</tbody>
</table>

Table 4.7: ‘Pass’ vs ‘fail’ rates with former and new cut scores applied

A breakdown of this analysis by grade level shows that 34 candidates received an OET A, 90 a B, 64 a C and 12 a D. As noted in the literature review chapter, a B grade is the minimum passing standard for the majority of health authorities using the OET. If the new cut scores were applied to the 200 test candidates’ writing samples, 32 more A’s rather than B’s would have been achieved. Conversely, 39 more test candidates would

\[ N = 200 \]

<table>
<thead>
<tr>
<th></th>
<th>With former OET cut scores applied</th>
<th>With new cut scores applied</th>
</tr>
</thead>
<tbody>
<tr>
<td>A + B = Pass</td>
<td>124 (62%)</td>
<td>85 (42%)</td>
</tr>
<tr>
<td>C + D = Fail</td>
<td>76 (38%)</td>
<td>115 (58%)</td>
</tr>
</tbody>
</table>
have failed to reach the minimum B grade requirement receiving a C rather than a B. Also, 56 more D’s rather than C’s would have been awarded. In other words, with the original OET-supplied test candidate cut scores, there were 124 A’s and B’s or ‘pass’ grades overall (62%) and 76 C’s and D’s or ‘fail’ grades overall (38%). However, when the new cut scores were applied, this changed to 85 A’s and B’s or ‘pass’ grades overall (42%) and 115 C’s and D’s or fail grades overall (58%) (see table 4.7). This is a significantly higher ‘fail’ rate.

4.4.3 Summary

To sum up, the quantitative analysis was undertaken as part of the examination of research question two (i.e. Is there any variability between judges in what they attend to while setting standards?) In addition, research question two also aligned (in part) to the assumption from Knoch and Macqueen’s (in preparation) argument-based framework “variation in qualitative and quantitative judgements of standard-setting panellists is within expectations (or can otherwise be controlled)”. Variability was attended to in terms of severity/leniency and consistency through the FACETS (Linacre, 2017) analysis and outlier judges were removed before the final cut score calculation was determined. Furthermore, the new cut scores that were generated from the analysis described above constitute a response to research question four (i.e. What occupational specific minimum standards (cut scores) do health professionals set on the Occupational English Test (OET) Writing sub-test?) The new scores that were derived from input from domain experts (doctors) were significantly closer in range to each other than the original OET cut scores. In addition, the application of these new scores would have dramatically affected the ‘fail’ rates for test candidates in the present data set and presumably, for other candidates taking the OET. The implications of these new cut scores and validity issues are further explored in the following discussion chapter.
4.5 Chapter summary

This results chapter firstly presented the qualitative findings and several themes that were derived from the thematic analysis. The nine main themes were: ‘Performance Level’ (PL) judgement, ‘Task Fulfilment’ (TF), ‘Content’ (C), ‘Organisation’ (O), ‘Expression’ (E), ‘Presentation (PRES)’, ‘Professionalism’ (PRO), ‘Audience Recognition’ (AR) and ‘Other’ (OTH). In addition, the topics of participants’ performance level judgements and decision-making and performance level judgements and variability was remarked on. The next section took note of panellists’ final evaluation of the standard-setting procedure (Training, Final and the Think-aloud (TAP)). The final section introduced the quantitative results which described the FACETS (Linacre, 2017) analysis and how rater severity/leniency and consistency was managed. Also, the new cut scores generated in this study and their possible implications were examined.
Chapter 5: Discussion

5.1 Introduction

The following discussion chapter is framed around the question of whether domain experts can set valid standards in a specific-purpose language (LSP) test. As noted in the literature review, the OET Writing sub-test and its status as an LSP test, with tasks that simulate requirements of real-world settings, is intended as a language assessment (rather than test of clinical competence). This separation of language and clinical competence is an Australian legal requirement (Jacoby & McNamara, 1999). It was argued earlier in this thesis that the judgements of domain experts (as end users/receivers of NESB test candidates' written communication in the workplace) are worth attending to as they are the ultimate arbiters of whether workplace communication is successful and may be more directly affected by the test outcomes than most stakeholders (see Buckendahl, 2005; Elder et al., 2017; Kenyon & Römhild, 2014; Manias & McNamara, 2016; Pill & McNamara, 2016). However, while a subject-matter expert perspective is arguably important, can domain experts without language expertise set valid standards on a language test? How do we know that their judgements are grounded in the construct of communicative competence as embodied in the OET? Can doctors separate issues of language and communication from those of clinical competence that they are clearly qualified to assess? These considerations are pertinent to the validity of their judgements and the resultant standards that are set on the OET Writing sub-test.

As stated in the literature review, the inclusion of domain experts in standard setting is somewhat taken for granted in LSP contexts and few studies have investigated the validity of their judgements in any depth (Kenyon & Römhild, 2014). Moreover, there is an assumption that the assessments of participants setting the standard are drawn from the same underlying construct and parallel the criteria used to assess performance. However, the extent to which this is true remains unclear and deserves to be further explored with qualitative standard-setting studies (Buckendahl, 2005).
Whether domain experts (rather than language experts) are better placed to set professional-ready standards on an LSP test is uncertain. This contention is further discussed in relation to the study’s findings.

This discussion chapter firstly summarises the themes from this study’s qualitative findings about what domain experts attended to while setting standards. This then leads to a visual presentation or model (see figure 5.1) of what participants addressed in the standard-setting process. Secondly, the elements from the qualitative results which are considered construct relevant or irrelevant to the current OET Writing sub-test construct, as operationalized through the test criteria, are illustrated. An explicit link is highlighted between the qualitative data and the validity question posed above. This illuminates the enquiry into the construct underlying domain experts’ judgements and the extent to which these are language-based. As part of the exploration of the validity of domain experts’ judgements the question of whether these expert judges can agree on appropriate standards is also considered, since the question of reliability is central to any validity argument. Variation in the qualitative and quantitative data and its impact on the stability of final performance standards and cut scores is therefore discussed.

Following this, domain expert feedback from the overall final evaluations (post standard setting training workshop, think-aloud and final evaluation questionaries) is examined. Domain experts’ views on the standard-setting process can also contribute to supporting and qualifying any conclusions about how valid their judgements are. Furthermore, comments from domain expert participants related to the OET Writing sub-test’s authenticity to the real world is considered as issues of task authenticity are pertinent to standard setting validity as well. Also, the legislative requirement that the OET test confines itself to (i.e. measuring language proficiency only) and its implication is deliberated. A final conclusion is then drawn about the wisdom of involving domain experts in standard setting. Individual variability emerging from both the qualitative and quantitative data analysis (and how it was dealt with) is taken into account along with the contribution of domain expert feedback on the standard-setting procedure. This leads to a discussion about the results of the standard-setting workshops and the weight that should be attached to the new cut scores derived from the standard-
setting process. Since these new cut scores are substantially different from the ones currently used for reporting purposes, careful consideration needs to be given to the defensibility of using them. The issues mentioned above are all pertinent to the assumptions associated with the decisions inference in the interpretive argument for LSP test validation laid out in the literature review and elaborated in the methods chapter in relation to the OET Writing sub-test. The final conclusions drawn from the evidence gathered in the implementation of this argument-based framework (Knoch & Macqueen, in preparation) are summarised in the conclusion chapter.

5.2 Domain expert decision-making process during standard setting

The qualitative findings presented in the results chapter outlined the themes and values that domain experts attended to when making performance level judgements, thereby offering insights into the relevance of these judgements to the OET Writing sub-test construct. The nine main themes identified were: ‘Performance Level’ (PL) judgement, ‘Task Fulfilment’ (TF), ‘Content’ (C), ‘Organisation’ (O), ‘Expression’ (E) ‘Presentation’ (PRES), ‘Professionalism’ (PRO), ‘Audience Recognition’ (AR) and ‘Other’ (OTH). Eight of the nine main themes were further sub-divided into sub-themes, (as noted in results section 4.2 and in table 5.1 shown below).

<table>
<thead>
<tr>
<th>Theme</th>
<th>Sub-theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘Task Fulfilment’ (TF)</td>
<td>‘Quality’ (Q) ‘Clarity of Meaning’ (M)</td>
</tr>
<tr>
<td>‘Content’ (C)</td>
<td>‘Purpose’ (PURP) ‘Case Note Content’ (CNC) ‘Patient Identification’ (PID)</td>
</tr>
<tr>
<td>‘Expression’ (E)</td>
<td>‘Language’ (L) ‘Conciseness’ (CON) ‘Professional Tone’ (PT)</td>
</tr>
<tr>
<td>‘Presentation’ (PRES)</td>
<td>‘Layout’ (LO) ‘Text Qualities’ (TQ)</td>
</tr>
<tr>
<td>‘Professionalism’ (PRO)</td>
<td>‘Clinical Competency’ (CC) ‘Patient Awareness’ (PA)</td>
</tr>
<tr>
<td>‘Audience Recognition’ (AR)</td>
<td>‘Audience Awareness’ (AA) ‘Effort’ (EF)</td>
</tr>
</tbody>
</table>
Table 5.1: Themes and sub-themes valued by domain experts

The identification of these themes has allowed for the development of a model (see figure 5.1) that draws together the various factors that come into play when domain experts (doctors) are making their standard-setting judgements. The model is a visual presentation of what doctors attended to while setting standards in this study. This model may help conclusions to be drawn about the construct relevance and hence the validity of the domain experts’ standards. In setting OET Writing sub-test standards, the doctors drew on and firstly used as ‘Input’ the OET Writing task case notes. As stated, these case notes are intended to replicate a real-world clinical situation and form the basis of a test candidate’s writing response. Secondly, as input, doctors utilized their own professional/clinical knowledge and experience to judge test candidates’ writing samples. Subsequently, doctors in an ‘Interpretation’ stage, focused on the particular ‘Textual Features’ evident in test candidates’ writing responses. These textual features are characterized under the thematic headings presented in table 5.1 above. What appears to be a fundamental overarching aspect of the textual features doctors commented on was the notion of ‘Audience Recognition’ (AR). That is, in the domain experts’ view, the letter writer (test candidate) must demonstrate an appropriate awareness of the recipient of the letter (another doctor) and how the letter will be received by that person. Furthermore, the interpretation and perception of a ‘Test Candidate’s Background’ (TCB) is also a part of this stage of the standard-setting process. From the doctors’ comments it seems that the background of a candidate (i.e. whether the test candidate is local vs foreign, experienced vs novice or English as a first language vs English as a second language (ESL) speaker/writer) has an impact on their final judgements. These aspects are used in combination in the decision-making process when doctors make their overall ‘Performance Level’ (PL) judgements in the final ‘Output’ phase of setting standards.
5.3 Construct relevance of domain experts’ judgements

As stated in the literature review, there is an implicit assumption in standard setting that the interpretations of participants making judgements parallel the criteria used to evaluate the performance. In other words, the criteria of panellists match those of the underlying construct of the test under consideration (Kenyon & Römhild, 2014). However, this assumption deserves to be further investigated in LSP contexts as what language experts (as test designers/raters) and domain experts (as end users) value, may or may not always align, raising doubts about the construct relevance and therefore the validity of the domain experts’ decisions about whether candidates have reached a particular standard (see Pill & McNamara, 2016; Manias & McNamara, 2016). Therefore, the construct relevance of the themes identified in the study’s
qualitative results, and summarized in figure 5.1 above, are considered in comparison to the OET’s Writing sub-test’s assessment criteria.

5.3.1 Construct relevance

Medical domain experts often attended to the same language aspects as set out in the OET Writing sub-test rating scale rubric. From the qualitative results, many examples were evident of the doctors’ comments that are clearly construct-relevant in the sense that they relate to the criteria that are currently used to assess OET Writing performance. As noted in the literature review, the OET Writing sub-test presently uses five criteria: ‘Overall Task Fulfilment’; ‘Appropriateness of Language’; ‘Comprehension of Stimulus’; ‘Control of Linguistic Features (Grammar and Cohesion)’; and ‘Control of Presentation Features (Spelling, Punctuation and Layout)’. The writing responses are rated by language experts against these five criteria with each measure having six levels, 1-6 (level 6 indicates the highest response) (OET, 2017). The Cambridge Boxhill Language Assessment’s full rating scale is confidential and therefore cannot be divulged in this thesis. What matters to domain experts (doctors) and how their own values are related to the construct of the OET Writing sub-test is illustrated in this section.

The first criterion of the OET Writing sub-test is ‘Overall Task Fulfilment’. This is a global criterion which includes whether the test candidate’s letter is of the required length. This criterion is relevant to the theme ‘Task Fulfilment’ (TF) identified in the domain experts’ commentary and also the sub-theme ‘Quality’ (Q). Standard-setting participants frequently made references to the global quality of test candidate letters (coded as overall ‘Quality’ (Q) (summative)), however this was not as specific as a final ‘Performance Level’ (PL) judgement.

The second criterion of the OET is ‘Appropriateness of Language’ which covers the use of appropriate vocabulary and tone in the candidate response and whether the response is organized appropriately. Similar features of writing were invoked by the domain experts (i.e. the comments categorized under the theme of ‘Expression’ (E)
and the three sub-themes of ‘Language’ (L), ‘Conciseness’ (CON) and ‘Professional Tone’ (PT). Also, as noted in the results chapter, comments categorized under ‘Clarity of Meaning’ (M), a sub-theme of ‘Task Fulfilment’ (TF), were often linked to test candidates’ use of ‘Expression’ (E). The features included in the OET criterion of ‘Appropriateness of Language’ were also reflected in doctors’ comments about the organization of the writing scripts, coded by the researcher as ‘Organisation’ (O) with the associated sub-themes of ‘Discourse Structure’ (DS) and ‘Prioritizing’ (PR).

The third criterion of the OET Writing sub-test is ‘Comprehension of Stimulus’ which takes account of whether the candidate response shows the task’s case note content and context has been understood and whether relevant rather than unnecessary content information has been provided to the reader. This OET criterion accords with what the domain experts reported about the content and purpose of the letter which was coded by the researcher as ‘Content’ (C), with the related sub-themes of ‘Purpose’ (PURP) and ‘Case Note Content’ (CNC).

The next criterion of the OET is ‘Control of Linguistic Features (Grammar and Cohesion)’ which invites consideration of how effectively grammatical structures and cohesive devices of English have been used. The medical domain experts also paid attention to these features as indicated in the researcher’s codes ‘Expression’ (E) and in particular the sub-theme of ‘Language’ (L) which encompassed ‘Grammar’ (G). Regarding the use of cohesive devices in the OET criteria, this aspect was also considered important by doctors and was encapsulated by the researcher under the sub-theme of ‘Language’ (L) and the codes ‘Language General’ (LG) and ‘Vocabulary’ (V).

The final criterion of the OET Writing sub-test, ‘Control of Presentation Features (Spelling, Punctuation and Layout)’, invites judgements on how these features affect the message that has been communicated. Doctors’ judgements were also influenced by spelling and punctuation, as evidenced in the comments coded under the theme of ‘Expression’ (E) and in particular the sub-theme of ‘Language’ (L) which incorporated the code of ‘Spelling’ (SP). Furthermore, the study’s theme of ‘Presentation’ (PRES) included the sub-themes ‘Text Qualities’ (TQ) and ‘Layout’ (LO).
The sub-theme ‘Text Qualities’ (TQ) took in the aspect of ‘Punctuation’ mentioned in the OET criteria. Also, this study’s theme of ‘Layout’ (LO) incorporated participants’ general comments about the use of standard formal letter layout, a concern which matches what is required by the OET criteria. As noted in the introduction chapter, the OET’s (2017) website states in regard to letter layout:

> A number of different formats are in accepted use by health professionals in different local contexts. There is therefore no single particular format that you have to use in your response in the OET Writing sub-test. It is important that your letter is clearly laid out and appropriate for the particular task, but there is no set OET layout that you have to use.

However, as noted in the results chapter, standard-setting doctors in this study made a clear distinction between the general layout of a test candidate’s letter and the issue of ‘Patient identification’ (PID) which was coded separately. Again, as stated in the results chapter, even though patient identification is assessed as an aspect of ‘Layout’ in the present OET criteria, because this feature was frequently referred to by standard-setting participants as an essential constituent of a referral letter, it was encompassed within the theme ‘Content’ (C).

In summary, when setting OET writing standards, doctors employ as ‘input’ the OET task case notes and their own professional/clinical knowledge to interpret the ‘textual features’ present in test candidates’ writing responses and make final performance level judgements or ‘output’ (see the standard-setting model and figure 5.1). There is reasonably close correspondence between these textual features and the aspects covered in the OET Writing sub-test criteria. Therefore, a preliminary conclusion can be drawn that domain experts’ judgements are largely consonant with the construct of communicative competence as embodied in the OET criteria. Nevertheless, there were some features that did not always match with the current OET Writing sub-test criteria and these features are discussed in the next section.

### 5.3.2 Construct disparity

Even though medical experts’ standard-setting judgements were often construct relevant, in the sense of being aligned with the present OET Writing sub-test criteria,
they also attended to features not currently covered by the OET criteria or gave greater weight to some aspects. The particular features they commented on, which extend beyond the OET construct as operationalized in the current assessment criteria, are discussed in the following sections.

5.3.2.1 Clinical competence

As shown in the above standard-setting model (figure 5.1), in setting OET Writing standards, doctors drew on their own understanding of the OET task case notes and their professional/clinical knowledge and experience to judge test candidates’ writing samples. This was in spite of the fact that, following the lead from Lumley’s (1998) study, standard-setting participants in the current study were told that they should not take into account a test candidate’s clinical knowledge while judging writing responses because, as noted, it is a legal and professional registration requirement of the Australian government that clinical knowledge and language proficiency be tested separately (Jacoby & McNamara, 1999; McNamara, 1996). As reported in the results chapter, the ‘Clinical Competency’ (CC) of a test candidate’s skills and knowledge, demonstrated through their interpretation of the OET Writing task’s case notes, was an issue raised by approximately two thirds of the panellists. For example, a test candidate may have been judged by a domain expert as having poor ‘Discourse Structure’ (DS) and limited ‘Prioritizing’ (PR) skills and/or a seeming lack of understanding of the ‘Case Note Content’ (CNC). Whether this perception actually influenced a standard-setting participants’ final ‘Performance Level’ judgement (PL) however remains uncertain. Also, as noted in the results chapter, many other construct relevant textual features and their related themes were noted more regularly and frequently by panellists and these may have had a more powerful influence on participants’ final performance level judgements.

The separation of communication from other aspects of clinical competence (noted as a concern of LSP testing by Douglas, 2001b and Jacoby & McNamara, 1999) was not always managed, as seen in the qualitative data examples in the results chapter. Qualitative data showed the blurring of clinical competency and language proficiency.
Therefore, perceptions of ‘Clinical Competency’ (CC) may have had an influence on a
standard-setting participant’s overall performance level categorization whether
consciously or unconsciously. The simulation of clinical situations allowed for
judgements of clinical competence, even when panellists knew they should be
excluded. This finding echoes that of Manias and McNamara (2016) who, in their
qualitative standard-setting study on the OET Speaking sub-test, showed that the
clinical scenario role-play simulations provoked judgements of clinical competence
from domain experts in spite of their being instructed to discount it. Jacoby and
McNamara (1999) state that communication and professional skills may be
“inextricably entwined” (p. 234) in an LSP test such as the OET and this was evidenced
to some extent in domain experts’ judgements in this study.

In summary, there were limits to the ‘commensurability’ between the OET Writing
sub-test criteria and what underlay the domain experts’ standards as their sense of
what mattered in the clinical context was somewhat “broader and differently
focused” (Pill & McNamara, 2016, p. 15). The fact that LSP tests involve simulations of
real-world performance may be a problem for standard setting in LSP testing in
general if they aim to measure language and communication skills alone, as Davies
(2001) would have it. Douglas (2000, 2005) on the other hand sees subject-specific
knowledge and skill as an essential component of the LSP test construct and would
presumably regard the participants’ concern with clinical competence while judging
language proficiency as both natural and appropriate.

5.3.2.2 Patient identification

It was established by this researcher, in correspondence with Cambridge Boxhill
Language Assessment, that ‘Patient Identification’ (PID) is included under the current
OET criterion of ‘Control of Presentation Features (Spelling, Punctuation and Layout)’.
Nonetheless, according to Cambridge Boxhill Language Assessment, this criterion is
not intended to be given more weight than other OET Writing sub-test criteria.
However, as noted in the results chapter, standard-setting participants valued this
feature highly seeing it as an essential constituent of the letter’s ‘Content’ (C) rather
than just a desirable feature of the letter layout. For some participants if patient ID (or elements of it) were not present their overall performance was deemed to be unsatisfactory. In making such decisions they may have been influenced by the Royal Australian College of General Practitioner’s (RACGP) (2016) standards for referral documents where a key criterion is that letters should “contain at least three approved patient identifiers” (p. 16) in the interests of facilitating optimal patient care in correspondence between health professionals. The fact that these real-world standards were applied directly to judging the quality of a simulated writing performance, without regard for the inevitable artificiality of the task, could be regarded as a test authenticity issue (Douglas, 2001b) in the sense that participants have, in this instance, treated the writing samples very literally as ‘real-world’ artefacts rather than as proxies for assessing writing competence. Alternatively, the categorical attitude of some participants could be seen simply as a responsible fulfillment of their gatekeeping role which includes ensuring that test candidates are minimally competent and can function safely and effectively in the workplace. This view would be consistent with the idea that the OET Writing sub-test criteria, as currently formulated, do not adequately reflect the priorities of the real-world setting (see further discussion below).

5.3.2.3 Audience recognition

As noted in results section 4.2.8, a quality that domain experts valued above all others when setting performance standards was sensitivity to audience or ‘Audience Recognition’ (AR) as it was classified by the researcher, encompassing the three sub-themes of ‘Audience Awareness’ (AA), ‘Effort’ (EF) and ‘Handoff’ (HO) (see also the standard-setting model, figure 5.1). The qualitative data from the training workshops and think-alouds showed that doctors in real-world settings who received a referral letter from another doctor were frequently ‘time-poor’ and worked in often busy and stressful environments. Hence, the awareness and recognition of the intended reader by the writer of the letter and their use of the identified ‘textual features’ was considered to be of upmost importance. This was somewhat at odds with the present OET Writing sub-test criteria which, rightly or wrongly, do not explicitly reflect this
notion of audience recognition and the value which was placed on it by all the standard-setting doctors in this study. However, as seen in the model (figure 5.1), it could be said that ‘Audience Recognition’ (AR), as an overarching category, was implicitly addressed in the textual features currently represented in the OET Writing sub-test criteria. This all-embracing aspect takes into account the textual features already noted by standard-setting participants as important when forming performance level judgements. It might also be argued that rather than characterizing domain experts’ concern with audience as somehow aberrant or construct-irrelevant, the OET Writing sub-test criteria should be refined to include more explicit mention of this feature. This would be in McNamara’s (1996) terms strengthening an LSP test from ‘weak’ towards ‘strong’ by “including for assessment further aspects of performance that are meaningful in the workplace context” (p. 56). The matter of possible revisions to the OET Writing sub-test criteria is discussed later.

In the meantime, it should be noted that some notion of audience awareness is central to all models of communicative competence, despite the variability between them (see e.g., Bachman, 1990; Bachman & Palmer, 1996, 2010; Canale, 1983; Canale & Swain, 1980; Hymes, 1972; Kramsch, 1986). Kramsch’s (1986) notion of ‘interactional competence’, regards performance and context as key where language use is a ‘co-construction’ (Jacoby & Ochs, 1995) and this is particularly relevant in the case of workplace language use. The conceptualization of the co-constructed engagement or awareness of the intended reader of the text (in this case referral letter) accommodates the view that the sender of the text needs to factor in how it will be received and interpreted. Viewing ‘Audience Recognition’ (AR) from the lens of Kramsch’s (1986) and Jacoby and Ochs’s (1995) interpretation of competence, participants engaged in interactions, whether spoken or written, need to be aware of their own socio-cultural norms and how they might be interpreted by others. This could include, for example, customs concerning the format, layout, content, organization and use of language in a formal document such as a referral letter which involves handover of information from one health professional to another.
5.3.2.4  Test candidate background

As noted in the literature review, IMGs are not a homogenous group, coming from a range of countries (including those where English is not the official language) and with a variety of medical experience (e.g. specialists or recent graduates) (Department of Health and Ageing, 2011; Hawthorne, 2012). As shown in the standard-setting model (see figure 5.1), consideration of a ‘Test Candidate’s Background’ (TCB) is a part of an interpretation stage of the standard-setting process undertaken by the doctor participants in this study. In examples noted in the qualitative results section 4.2.9, judges at times made reference to the background of a test candidate (i.e. local vs foreign, experienced vs novice or English as a first language vs English as a second language (ESL)) and indicated that these aspects may have had an impact on their final performance level judgement. As mentioned in results section 4.2.9, (see also Appendices K and L) half of the participants in this study (9 out of 18) had ‘moderate’ to ‘high’ workplace interaction with non-English-speaking background (NESB) colleagues, and furthermore half of all participants had moderate to high levels of experience reading referral letters written by a NESB doctor. However, the remaining participants mentioned they had limited interaction/reading experience with NESB doctors. In either instance, this may have had an impact (positive or negative), on a participant’s view of test-taker performance.

Regardless of their experience, domain experts in this study might have been susceptible to the negative perceptions of IMGs in Australia that emerged in reports into language proficiency standards in Australia such as ‘Lost in the Labyrinth’ (House of Representatives Standing Committee on Health and Ageing, 2012) and to anecdotal evidence that NESB OET test takers “were passing the test with inadequate proficiency in English to cope with the demands of their profession” (Lumley, 1998, p. 352). Indeed, as previously noted, even the definition of the term IMG is ‘problematic’ as it may lead to an interpretation that an IMG still has a “deficiency (e.g., in language, professional knowledge or cultural competence)” post-registration (Pill, 2013, p. 16). The very existence of studies of bias against IMGs in the workplace (see Elkin et al., 2012; Harding et al., 2010; Louis et al., 2010) indicates that there may be a continued
perception that IMGs might not be up to the same standard as a locally-trained medical professional.

Test candidate background (TCB) is not mentioned in the OET Writing sub-test criteria, which rightly focuses on the quality of the writing rather than the writer. For reasons of fairness to all test candidates, there is an expectation of equal treatment and any evidence of bias in panellists’ standard setting judgements should be of concern. However, as noted in results section 4.2.9, references to a ‘Test Candidate’s Background’ (TCB) were not made by all panellists (50%) and were restricted to a small number of test writing samples. Hence, the impact of this factor on participants’ final ‘Performance Level’ (PL) judgements may be negligible.

5.3.2.5 The OET Writing sub-test task and authenticity

A further issue addressed in the domain experts’ feedback concerned the OET Writing sub-test task itself, which clearly had a bearing on standard-setting judgements. As noted in results section 4.2.9, an area of some concern for medical expert standard-setting participants in this study was the authenticity of the OET Writing sub-test task. Roughly half of all participants pointed out that the task of writing a referral letter in the current OET format is one that does not exactly represent the present workplace reality for some doctors. The issue of tasks attempting to mirror real-world contexts and situations is not just a complex one for the OET, but for all LSP tests (Douglas, 2010; Kane, 2004, 2006). Manias and McNamara (2016) argue that some domain experts in their study found the task of standard setting on the Speaking component of the OET was complicated by the fact that the participants had reservations with the task (Manias & McNamara, 2016). In the current study, participants remarked that the current OET Writing sub-test task response requires that test candidates present the content of the letter in full paragraphs and in handwritten form, whereas, as mentioned in results section 4.2.9, the reality is that many doctors today write such letters, at least partially, in bullet or note form and they are rarely hand-written. This may be partly a function of time pressure, but is also related to ‘Audience Recognition’
 Nevertheless, as stated in results section 4.2.9, doctors still, to a degree, and perhaps in an emergency situation, hand-write letters and also in full paragraphs as in the current OET Writing sub-test task. A number of studies have shown that referral and reply letters are still an essential part of written information exchange between health professionals (Berendsen et al., 2009; McConnell et al., 1999; Piterman & Koritsas, 2005). Macqueen et al. (2012) noted in their study of stakeholder views on the OET that doctors found “the task of composing a formal handover letter, e.g., referral or discharge, from case notes was highly relevant”, “was closely linked to their daily practices” and considered that the OET task of “selecting, synthesising and transferring information into formal prose for receipt by another health professional to be worthwhile” (p. 18). Hence, the current test format and design may still largely be applicable to the real world. On balance then, despite the limitations to task authenticity that were mentioned by some standard-setting panellists it seems unlikely that the format of the task would constitute a major obstacle to valid decision-making.

In sum, bearing in mind the various factors underpinning domain experts’ judgements that have been discussed in the previous section it would seem reasonable to conclude that their decisions were partly based on factors relevant to the OET Writing sub-test construct, in the sense that they were often well aligned to the OET criteria as currently formulated. However, while there was reasonably close correspondence between the features covered in the OET criteria and the ‘Textual Features’ in the standard-setting model (see figure 5.1), the informants departed at times from the OET construct by paying heed to professional and clinical knowledge in some instances, by rewarding evidence of ‘Audience Recognition’ (AR) (although this is arguably already implicit in the current OET criteria) and perhaps also in drawing inferences about ‘Test Candidate Background’ (TCB) when making judgements about the quality of the written samples. To the extent that these features informed their decisions they could be interpreted as construct-irrelevant and hence as jeopardizing the validity of their judgements. On the other hand, it has been argued, in the case of
‘Audience Recognition’ (AR) and ‘Clinical Competence’ (CC), that the OET’s failure to accommodate these features could instead be seen as a limitation of the test rather than a problem with the panellists’ decision-making processes. The question of whether the OET could be revised to achieve better alignment with health professional values is briefly entertained in the conclusion chapter.

In the meantime, it is appropriate to turn to a further matter bearing on the validity of standard setting judgements, namely the extent of variation in domain expert participants’ decision making as evidenced in both the quantitative and qualitative data.

5.4 Variation in qualitative data

The use of domain experts in this standard-setting study inevitably resulted in variance between what each participant attended to when setting standards. This is because each individual judge has their own unique professional background/experience and perhaps utilized their own ‘indigenous assessment’ criteria (Jacoby & McNamara, 1999). This variability was evident in all aspects of the decision-making process (as represented in figure 5.1) including the extent to which the participants referred to the OET Writing task case notes, and whether and how much they drew on their own professional/clinical knowledge. There was also some inconsistency and variability in the way in which participants judged the ‘Textual Features’ shown in test candidates’ writing samples.

As demonstrated in results section 4.2.11 and the qualitative examples from the training workshops and think-aloud sessions, the main areas of dissonance between judges included the different degrees of emphasis given to the following overall themes identified by the researcher: ‘Patient Identification’ (PID), ‘Organisation’ (O), ‘Expression’ (E) and ‘Presentation’ (PRES) features. Some participants regarded these aspects dissimilarly. As stated in results section 4.2.11.1 a key instance of this dissonance was to do with ‘Content’ (C) and the inclusion or not of ‘Patient Identification’ (PID). In this example, the judge (J12) has indicated that the lack of
identification has meant that they significantly lowered their overall performance level judgement:

If he had of identified the patient I would’ve gone “between Strong and Competent” but he didn’t identify the patient so again I’m going to have to say “Not Yet Competent”. Look if he had of identified the patient I would’ve marked him up. I feel like I’m nit-picking but, I think it’s, this is clinical medicine. It has to be done properly. [J12, TAP – 226].

As stated, this was to some extent to be anticipated as judges were asked to utilize their own criteria when making final judgements which may or may not differ to the current OET Writing sub-test criteria. As noted in the literature review, while domain experts may have a better understanding of workplace language requirements and their judgements are likely to more closely mirror the language competency demands of those environments (see e.g., Douglas, 2001a; Elder, 2001; Hamp-Lyons & Lumley, 2001; Knoch, 2009; Knoch, 2013; Pill, 2013; Pill & McNamara, 2016; Manias & McNamara, 2016), “standard settings are based on the judgements of multiple individuals, so variability in judgements is an expected outcome” (Kenyon & Römhild, 2014, p. 5). As mentioned in the literature review, IMGs cannot be classed homogenously (Pill, 2013) and therefore likewise their locally trained counterparts cannot also be considered in the same way. Variability is all the more likely between individuals with somewhat diverse professional backgrounds and different amounts of training and experience as was the case in this study. As noted in the method chapter (and also in Appendices I, J, K, L and M), a range of participants with varied backgrounds is desirable in order to ensure that the findings can be regarded as generalizable as possible. However, it should also be acknowledged that the variability between participants’ focus of attention on the themes mentioned above could have a potential effect on the final performance standards and subsequently the validity of the cut scores. The impact of this variability, and how it was attended to and dealt with, is further discussed in the following section on quantitative variation.
5.5 Variation in quantitative data

As mentioned, variation may be a threat to the validity of cut scores when using domain experts. The question is whether it can be managed sufficiently to produce judgements that are acceptably reliable. As stated in results section 4.4, there was some variation between participants evident in the quantitative data and that is in the way in which domain experts classified the different writing samples. The many-facet Rasch program FACETS (Linacre, 2017) was used prior to final cut scores being determined. The analysis took into account possible aspects such as discrepancies in judges’ severity/leniency and consistency.

Regarding rater severity/leniency there was some variation between participants. One judge (J6) was especially severe (logit score of 1.79) and another (J2) was somewhat more lenient (logit score of -1.03) than the other panellists. Furthermore, the rater separation index figure of 4.07 showed that the variance between judges was approximately four times the error of estimates. If all participants were equally severe/lenient, the standard deviation of the judges’ difficulty estimations would be parallel or less than the mean estimation error of the entire data set (Bond & Fox, 2015). This demonstrated that the judges were not rating writing performances in a congruently severe/lenient manner (i.e. there was substantial difference between them). In addition, the reliability figure of 0.94 indicated the FACETS analysis was quite reliably separating participants into different levels of severity/leniency (a low reliability figure indicates judges as equally severe/lenient). In sum, the separation index and the reliability measures showed that the standard-setting participants were not judging writing samples in an equally severe/lenient manner – a level of variability is desired to some extent across a full range of performance levels.

There was also some variation regarding each judge’s rating consistency. One judge (J6) was assigning performance level judgements somewhat unpredictably and exhibiting substantial misfit (Infit MnSq: 2.07). This is well above the tolerated upper range of 1.5 of Lunz and Stahl (1990) and Linacre (2017). Therefore, J6 appears to have been allocating performance level judgements inconsistently. Nevertheless, most
Panellists seemed to be making performance level judgements relatively consistently and across an acceptable range of performance level categories. As stated, the intent of the AJM is not for total and exact participant agreement and this is not to be anticipated for judges in general and especially for those allocating performance level judgements without stated criteria to guide them. However, as Linacre (2017) notes, some raters may need to be removed from the analysis (i.e. those who show idiosyncratic tendencies which could have an adverse effect on the overall results). Hence, J6’s inconsistent rating behaviour further confirmed his removal from the final AJM cut score calculation.

Overall, the evident quantitative variation reported in the results chapter was within reasonable limits and could therefore be attended to and managed statistically. The analysis showed that the majority of participants were indeed rating the writing samples and making final performance level judgements consistently and with any variation in severity falling within an acceptable range. Also, there was a tolerable level of difference and consistency in their performance level allocations. On the whole, the data fit the model adequately and variability was at a manageable level. In this respect, therefore, it can be concluded that domain expert judgements were consistent enough to be regarded as trustworthy.

5.6 Domain expert feedback

In addition to the qualitative and quantitative findings, the domain expert feedback further confirmed that even though there was some minor variation between participants, most believed that they were able to make performance level judgements successfully. As noted in results section 4.3, participants were asked for their feedback on the standard-setting process as part of an overall final evaluation. This was undertaken at the end of the training workshops, think-aloud sessions and a final evaluation. The majority of panellists stated that they understood the AJM method’s performance levels, their differences and how to make judgements. Also, all understood the standard-setting task and stated that the training was sufficient to complete it effectively. Therefore, overall, the domain expert feedback on the
standard-setting sessions generally endorsed the view that the participants were comfortable and confident in their decision and judgement-making ability.

Furthermore, the participant feedback given in the final evaluation showed that all believed the standard-setting process and experience was valuable to them and important. Most also agreed that further standard-setting procedures should be carried out regularly in future. Some stated that standard setting is necessary in terms of consistency and fairness for test candidates and also in terms of patient safety and connections to real-world medical practices. These findings offer further support for the argument that domain experts should be included in the standard-setting process and can make valid judgements.

The discussion in this section has spoken to the question of whether domain experts should be used in the standard-setting procedure of an LSP test such as the OET. The themes illuminated through the analysis of the doctor panellists’ commentary, revealed in this instance, a reasonable degree of alignment between the domain experts’ views of what was important for writing handover documents in a workplace setting and the criteria embodied in the OET Writing task criteria. The standard-setting doctors in this study appeared to be attending to most aspects of performance that the test is designed to measure. Furthermore, variation in the quantitative data was dealt with and managed by removing outliers and an inconsistent judge before the final cut scores were calculated. Also, the participant feedback supported the stance that domain experts were able to adequately make standard-setting judgements and that they themselves viewed their contribution as integral to the process.

A counter view could be mounted to the effect that decision-making by some participants and with respect to some of the writing samples, may have been influenced, rightly or wrongly by factors that were extraneous to the OET Writing sub-test construct as currently operationalized. Support for this view comes from the evidence in the commentary from some of the doctor participants that aspects of clinical competence and missing content information such as details of patient ID may be overriding judgement of the communicative adequacy of some of the writing samples. It could be argued, however, that this was not apparently true for most of
the participants and that in any case maverick judgements were identified and
removed from consideration. It also seems reasonable to claim, that given the current
constraints of the OET, this pragmatic approach to dealing with possible construct
irrelevant variance, is the best solution to the issues identified, at least until such time
as the OET Writing task and the limits to its authenticity can be addressed. Subject-
matter expert involvement in the standard-setting procedure is arguably better than
no such involvement and the weight of evidence supports taking this expert
judgement seriously. The following section therefore discusses the new cut scores
that were informed by the judgements of the domain expert participants.

5.7 New cut scores

The cut scores generated from domain expert input are discussed in this section. As
noted in the quantitative results section, the new cut scores are very different from
the current ones, and should probably be taken seriously by Cambridge Boxhill
Language Assessment in light of the arguments formulated above. The results of the
quantitative analysis showed that the new cut scores are generally more stringent
than the current ones. As stated in results section 4.4.2, for the Analytic Judgement
method (AJM) performance level ‘between COMPETENT and NOT YET COMPETENT’
(the minimum needed for an OET B grade or ‘pass’) the new cut score was somewhat
higher (and therefore more stringent) than the present OET. Also, for the AJM
performance level ‘between NOT YET COMPETENT and UNSATISFACTORY’ (or the
minimum requirement for an OET C grade) was significantly higher (i.e. harder to
attain) than the current OET. Furthermore, the new cut score for ‘between STRONG
and COMPETENT’ (the minimum score required for an OET A grade) was slightly less
stringent than the present OET.

The validity of cut scores is also buttressed by the steps taken to ensure that a suitable
method was chosen (i.e. AJM) compared to other possible standard-setting
procedures, as noted in the literature review and method chapter. Consistency in the
running of the workshops and think-aloud sessions used in this study was confirmed
by the researcher as attempts were made to ensure that they were conducted in a
similar manner and the same materials were used throughout. This verifies the accuracy and consistency of the cut score results. In addition, suitability of the AJM standard-setting method was corroborated, in part, through comparison with two other recent standard-setting studies that also utilized the AJM and showed comparable cut score results. Cizek (2012) notes that the “degree of heterogeneity among applications of any one method is remarkable” and that “the characteristics of standard-setting procedures for a specific application vary from the characteristics of the same procedure applied in other contexts, it becomes increasingly difficult to conduct focused research on a method, or even to operationalize the distinctive characteristics of the method one hopes to study” (p. 11).

However, further support for the validity of the cuts scores emerged through comparison with two other standard-setting studies. The recent application of the AJM in three recent standard-setting studies (including this one) showed that the method could indeed be utilized consistently. The trend displayed in the cut score results in this study is quite similar to that reported in Pill and McNamara’s (2016) standard-setting study on the OET Speaking sub-test. That study set new cut scores for three professions (medicine, nursing and physiotherapy). The results were remarkably comparable to the current study with that study’s OET equivalent B/C cut scores (pass/fail) being more stringent and the OET A/B cut score being more lenient, as was the case in this study. Another recent ARC-funded standard-setting study on the OET Writing sub-test for the profession of nursing (Knoch et al., 2017) again showed similar results to the present study with both the OET equivalent B/C (pass/fail) and OET C/D cut scores being more stringent and the OET A/B cut score being more lenient. Overall, in three separate recent standard-setting studies on the OET (including this study) that were informed by domain expert input, the new cut scores were generally more stringent than the current OET ones.

5.8 Chapter summary

This discussion chapter has offered insight into the question of whether the new OET Writing sub-test cut scores generated from this study are sufficiently valid and
dependable to be used in favour of the current ones. To this question, the study’s methodology, findings and discussion suggest an affirmative answer, albeit tentative. As noted in the above discussion, firstly the domain expert participants generally did attend to construct relevant features in the sense that their judgements were on the whole related to the different components of the OET Writing sub-test criteria. Secondly, in spite of the observed instances of variation in judgements emerging from the quantitative and qualitative analyses, steps were taken using the FACETS analysis to contain and compensate for the variability in the data. In addition, the feedback from the overall final evaluations of the standard-setting process indicated that the domain experts felt able to satisfactorily make standard-setting judgements despite their reservations about the authenticity of the OET Writing task. Finally, the procedures adopted for standard setting were rigorous and sound enough to generate defensible cut scores.
Chapter 6: Conclusion

6.1 Introduction

This conclusion chapter begins with a summary of the findings which are framed around Knoch and Macqueen’s (in preparation) argument-based approach to validity for an LSP test. Following this, the implications of the study for the OET Writing sub-test are presented. Next, the contribution of this study to standard-setting research is offered. The study’s limitations are then noted and areas for future research are outlined.

6.2 Summary of the findings around an argument-based approach to LSP test validity

Validity evidence for an LSP test like the OET Writing sub-test needs to be gathered if claims about test scores and the inferences drawn from them can be deemed defensible. As noted in the literature review, based on Kane’s (2004) argument-based approach, Knoch and Macqueen (in preparation) pose an overall validity argument for LSP tests, including the assumptions underpinning the ‘decisions’ inference which is of specific concern to standard-setting in general and to the current study in particular. These assumptions, as articulated in Knoch and Macqueen’s validity framework, help shape and summarize the findings in this study.

In the table 5.2 below, the warrant for the decisions inference is shown in the first column: “Estimates resulting from the performance on the LSP tasks are useful for decision-making about readiness (or similar purposes) for work in the TLU domain” (Knoch & Macqueen, in preparation). In the second column, the associated assumptions and related research questions from this study are given. Sources of backing actually collected for this study are provided in the third column. Confirmation of backing from this study (i.e. strong or partial with qualification) is presented in the fourth column.
**Decisions inference**: assumes that the decisions made based on the LSP test are appropriate and equitable

**Claim**: decisions made based on the estimates of the quality of the performance are appropriate and well communicated

<table>
<thead>
<tr>
<th>Warrant</th>
<th>Assumptions</th>
<th>Sources for backing</th>
<th>Confirmation of backing from this study</th>
</tr>
</thead>
</table>
| Estimates resulting from the performance on the LSP tasks are useful for decision-making about readiness (or similar purposes) for work in the TLU domain | The standards set on the LSP test reflect the language standards operating in the TLU domain. Relevant to research question 1 (i.e. What criteria do standard-setting participants use as a basis for their decisions in judging writing responses and to what extent are these decisions language-based?) | Involvement of domain experts in standard setting  
*In this study:*
- Qualitative comments provided by 18 domain expert participants from a range of health professions: general practitioners (GPs), specialists, consultants and medical educators | PARTIAL  
Standards mostly reflect workplace demands/requirements as interpreted by domain experts  
**BUT**  
Small sample of domain experts and OET samples may limit representativeness of standards set |
| Standard-setting panellists are oriented to construct-relevant features of the performance | Verbal protocols during standard setting; discussions during workshops  
*In this study:*
- Qualitative data from 7 training workshops and 5 verbal protocols (think-alouds – TAPs) sessions | PARTIAL  
Comments reflected OET criteria in the main  
**BUT**  
Decisions sometimes influenced by additional considerations. Construct relevance of these additional features is debatable |
| Variation in qualitative and quantitative judgements of standard-setting | Statistical analysis of score data; analysis of verbal protocol/discussion data | PARTIAL  
Observed quantitative variation within acceptable limits and could be managed |
| Panellists is within expectations (or can otherwise be controlled) | **In this study:** | statistically. Tolerable level of variance
| Relevant to research question 2 (i.e. Is there any variability between judges in what they attend to while setting standards?) | - Analysis of AJM cut score data and subsequent use of many-facet Rasch program FACETS (Linacre, 2017) | BUT
| | - Thematic analysis of verbal protocol/discussion data using NVivo software | Thematic analysis revealed decisions were sometimes influenced by additional considerations. Construct relevance of these additional features is debatable
| There is sufficient and appropriate information for the standard-setting panellists to make a decision | Interviews, surveys with standard-setting panellists | PARTIAL
| Relevant to research question 3 (i.e. How do standard-setting judges view the process and outcome of the standard-setting procedure?) | In this study: | OET Writing task generally perceived as providing sufficient and appropriate information
| | - Qualitative commentary from 7 training workshops and 5 verbal protocols (think-alouds – TAPs) sessions | BUT
| | Standard-setting panellists are confident in the validity of their judgements | Feedback from standard-setting panellists | STRONG
| Relevant to research question 3 (i.e. How do standard-setting judges view the process and outcome of the standard-setting procedure?) | In this study: | Final evaluation data and think-aloud data showed majority of panellists understood AJM method’s performance levels, their differences and how to make judgements. Also, all understood standard-setting task and stated training was sufficient to complete it effectively
| | - All 18 participants completed a training and final evaluation form (see Appendices Q and R). 5 panellists involved in think-aloud protocol (TAP) stage completed a think-aloud evaluation form (see Appendix S) | |
Standard-setting procedures are suitable and consistently applied

Relevant to research question 3 (i.e. How do standard-setting judges view the process and outcome of the standard-setting procedure?)

Justification for choice of method; detailed account of procedures; feedback from standard-setting panellists; comparisons with other standard-setting studies using the same method

In this study:
- Potential standard-setting methods outlined in the literature review and a comparative pilot study conducted on two possible methods
- Detailed account of AJM procedures given in the method chapter
- All 18 participants completed a training and final evaluation form (see Appendices Q and R). 5 panellists involved in think-aloud protocol (TAP) stage completed a think-aloud evaluation form (see Appendix S)
- Comparisons made with other recent standard-setting studies on use of AJM method with the OET (Knoch et al., 2017; Pill & McNamara, 2016)

STRONG
Two possible standard-setting methods justified and compared in a pilot study to guide final choice. Procedural refinements made to final choice (AJM)

Detailed account of AJM procedures given in method chapter. AJM consistently applied in all 7 workshops and 5 think-aloud sessions

Three separate recent standard-setting studies on OET (including this study) that were informed by domain expert input which used the same AJM yielded similar results (i.e. new cut scores were commonly more stringent than current OET ones)

<table>
<thead>
<tr>
<th>Table 5.2: Argument-based validity framework and backing from this study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knoch and Macqueen’s (in preparation) argument-based validity framework for an LSP test has been applied to the OET Writing sub-test. Examples of evidence of backing (presented in table 5.2 above) are discussed in relation to each assumption and relevant research question from this study.</td>
</tr>
<tr>
<td>The first assumption is that “the standards set on the LSP test reflect the language standards in the TLU domain”. It relates directly to research question one from this</td>
</tr>
</tbody>
</table>
study (i.e. What criteria do standard-setting participants use as a basis for their decisions in judging writing responses and to what extent are these decisions language-based?). This study’s sources of evidence to back this assumption included the input of 18 subject-matter experts from a variety of health professions such as general practitioners (GPs), specialists, consultants and medical educators. Evidence from these domain expert sources partially confirms this assumption allowing the claim to be made that the standards generally mirror the workplace demands/requirements as construed by subject-matter experts. Although the number of domain experts in the sample accords with what is considered acceptable practice (Zieky et al., 2008), the local nature of the sample drawn on the basis of convenience may nevertheless limit the representativeness of the standards to some extent given that the OET is used across multiple jurisdictions beyond Melbourne, Victoria where the current study was conducted.

A second assumption, “standard-setting panellists are oriented to construct-relevant features of the performance”, also relates to research question one from this study. This study’s source of evidence to back this assumption is the commentary from the seven training workshops and five verbal protocols (think-aloud) sessions. Concrete evidence of backing from this study partly corroborates this assumption because the features mentioned by domain experts when making their standard-setting decisions on the whole matched the features covered by the present OET Writing sub-test criteria, which can be seen as representing the current OET construct. Nevertheless, participants’ decisions were occasionally influenced by further considerations that were not related to the OET criteria and the construct relevance of these additional aspects, as discussed in the previous chapter, is debatable.

A third assumption is that “variation in qualitative and quantitative judgements of standard-setting panellists is within expectations (or can otherwise be controlled)”. This assumption relates to research question two from this study (i.e. Is there any variability between judges in what they attend to while setting standards?). Evidence backing this assumption comes from an analysis of the AJM cut score data and the successive use of many-facet Rasch program FACETS (Linacre, 2017) to account for any variation in domain experts’ judgements. In addition, a thematic analysis of the
verbal protocol and workshop discussion data using NVivo software was undertaken to compare any variation in domain experts’ decision-making processes. Findings from this study to a large extent confirmed this assumption as the observed quantitative variation was inside tolerable limits and could be managed statistically (i.e. there was an acceptable level of individual variance in aspects attended to amongst participants). However, the thematic analysis revealed that some participants showed a level of inconsistency in their decision-making perhaps because they were sometimes influenced by additional considerations which arguably fell outside of the construct represented by the OET Writing sub-test criteria.

The next assumption is: “There is sufficient and appropriate information for the standard-setting panellists to make a decision”. This relates to research question three from this study (i.e. How do standard-setting judges view the process and outcome of the standard-setting procedure?). As noted, this study successfully elicited commentary from seven training workshops and five verbal protocol (think-aloud) sessions as backing for this assumption. Findings revealed that the OET Writing sub-test task was seen by participants as mainly offering adequate information as a basis for making decisions about readiness for the workforce. Yet, a small number of domain experts stated that there were limitations in the task's authenticity and the features of the task that did not align to the corresponding writing requirements of the workplace made it difficult for them to judge readiness in some cases. In addition, a small number of judges expressed, rightly or wrongly, that they would have liked further direction from the workshop moderator or been provided with pre-established standards (i.e. pass/fail examples) to make performance judgements more effectively.

A further assumption (also related to research question three from this study) is that “standard-setting panellists are confident in the validity of their judgements” and backing for this assumption was sought via “feedback from standard-setting panellists”. In this study, every panellist filled in a training and final evaluation form. In addition, five participants who completed the think-aloud protocol (TAP) phase also filled in a think-aloud evaluation form. Findings from these sources strongly confirm this assumption. From the final evaluation of the standard-setting procedure, this
study’s evidence indicated that most panellists understood the AJM’s performance levels and the process of recording their judgements using designated forms. Furthermore, all participants confirmed that they understood the standard-setting task and affirmed that the training was appropriate to complete the task successfully.

A final assumption (also related to research question three) is that “standard-setting procedures are suitable and consistently applied”. Backing for this assumption came from various sources, including the pilot study which was used to weigh the benefits of using two possible standard-setting methods to introduce procedural modifications to refine the final choice (AJM). The comprehensive account of the AJM procedures provided in the method chapter can also be taken as evidence of the rigour and consistency with which these procedures were applied in all seven workshops and in the five think-aloud sessions. Final post-workshop and think-aloud evaluations demonstrated that all participants understood the AJM’s performance levels, their differences and how to make judgements. Likewise, the majority of panellists mentioned that they understood the standard-setting task, specified the training was adequate to complete it successfully and gave general endorsement of the study’s standard-setting process. In addition, comparisons were made with two other related standard-setting studies on the OET using the same AJM method (Knoch et al., 2017; Pill & McNamara, 2016). Comparative analysis partially backs this final assumption. The AJM was successfully applied in parallel standard-setting studies on the OET (i.e. Knoch et al., 2017 – OET Writing sub-test samples used with nurses and Pill & McNamara, 2016 – OET Speaking sub-test samples used with doctors, nurses and physiotherapists) and this could be seen as offering further support for this assumption. What is unclear, however, is whether the comparable results (i.e. more severe cut scores in each case) are an artefact of this method, or instead a reflection of more accurate and defensible standards.

In summary, the evidence supporting the overall validity of the standard-setting process and its outcomes is relevant to the central question discussed in the previous chapter of whether domain experts’ involvement in setting standards is appropriate and whether the cut scores derived from this process are defensible. The answer would appear to be yes in the current case (with certain reservations) and in general
provided that the assumptions underpinning domain expert involvement in the decision-making process can be supported.

6.3 Implications

The findings of this study’s standard setting procedure, with domain expert input, have practical and operational implications for the OET. The first to be considered is that the new cut scores were overall more stringent than the ones currently being used. If the new cut scores are to be employed operationally, Cambridge Boxhill Language Assessment will need to advise stakeholders and relevant medical boards about their potential effect on professional registration (i.e. the effects of the new cut scores on overall pass/fail rates). Agreement about the suitability of the new cut scores will need to be sought from regulatory authorities in the different jurisdictions served by the OET. Moreover, if the new cut scores are adopted, test candidates themselves will also need to be advised of the changes. These factors may have certain professional and commercial implications for Cambridge Boxhill Language Assessment and influence the views of the test itself by its stakeholders and end users.

Another practical concern relates to future standard-setting workshops being conducted and the matter of referral letter layout and inclusion of patient identification. One of the issues that was raised by the standard-setting doctors was around the inclusion or not of ‘Patient Identification’ (PID) as an important ‘Content’ (C) aspect of a referral letter. This feature is currently included under the OET criterion rubric ‘Control of Presentation Features (Spelling, Punctuation and Layout)’. As mentioned, the OET (2017) states on their website in relation to letter layout that “a number of different formats are accepted” and “there is therefore no single particular format”. Additionally, several sample letter layouts are given as models on the OET’s website which include, to varying degrees, the presence of features of identification such as a patient’s name, date of birth and address. As stated in the discussion section, this aspect was given far more weight and value by the doctor standard setters than in the current OET Writing sub-test criteria. Therefore, it is recommended that a benchmark letter format be employed by Cambridge Boxhill Language Assessment’s
test developers following the layout stipulated in The Royal Australian College of General Practitioners (RAGCP’s) (2016) standards. By having a standard letter format with the inclusion of appropriate patient ID, this would make this vitally important feature (as valued by domain experts) a standard part of all test candidates’ letters. A benchmark letter such as this would also likely neutralise the patient ID issue in further standard-setting sessions on the OET Writing sub-test for standard-setting participants. However, as the test is used beyond Australia, the same identical format may not be suitable in all jurisdictions in which the OET is accepted.

Furthermore, a standard letter format would also better prepare test candidates in training and preparation not only for the OET Writing sub-test, but also for the future workplace readiness. By taking this aspect into consideration, arguments that IMGs are not adequately equipped for the communicative requirements of work environments might be countered, at least in part. In addition, some of the standard-setting participants noted that additional direction would have been beneficial to specify how significant the presence and particularly the lack of patient ID had on their performance level judgements. In the absence of the revisions proposed above to the OET website’s letter samples, if future standard-setting sessions are conducted, the aspect of ‘Patient Identification’ (PID) should be considered by the workshop moderator and further brought to the attention of panellists (i.e. it should be highlighted and stressed to standard-setting participants that test candidates’ letters are produced in a simulated test environment and are not real-world documents).

An additional consideration is that perhaps the imposed Australian government legal requirement of separating language and clinical competence is somewhat inappropriate and not 100% achievable for domain experts in a standard-setting context given the view of Jacoby and McNamara (1999) that communication and professional skills might be “inextricably entwined” (p. 234). Also, Pill (2013) puts forward that “a language test with ‘traditional’ language assessment criteria may be ‘strengthened’, that is, moved along a cline from weak towards strong, in McNamara’s (1996) terms, by including for assessment further aspects of performance that are meaningful in the workplace context” (p. 56) (such as aspects noted earlier in this
study such as Clinical Competency (CC), Patient Identification (PID) and Audience Recognition (AR)). It should be noted that, independently of the current research which was conducted as part of a larger funded study, attempts have been made to revise the current OET assessment criteria similar to what Pill (2013) proposed, drawing on insights from health professionals. According to the project’s final report (Knoch et al., 2017), two key recommendations were made to Cambridge Boxhill Language Assessment: 1) the adoption of new professionally-relevant OET Writing sub-test criteria as informed from aspects of written communication valued by doctors, nurses and health information managers (i.e. indigenous assessment criteria – Jacoby, 1998; Jacoby & McNamara, 1999); 2) the revision of the existing specifications for the OET Writing task to contain situations and contextual information required to draw out the features of communication valued by participants, but presently not incorporated in the Writing task.

6.4 Contribution of the study

This study has made a number of relevant contributions to standard-setting research and particularly with regard to a language for specific purposes (LSP) test, namely the OET Writing sub-test. The first of these contributions is that this study has framed standard-setting in the context of a validity argument in line with recent thinking about the validation process. This has not been done so thoroughly before in language testing contexts and never to our knowledge for an LSP test. The second contribution is that the study offered a refined methodology for involving domain experts and validating their judgements, which could be applied in other LSP testing situations. The qualitative investigation of the basis for standard-setting judgements is an under-explored area in standard-setting research and should be seen as central in claiming the validity of resultant cut scores. This study also highlighted the age-old challenges for LSP testing and how they played out in the OET context. Even though the OET test is seen as an exemplar of an LSP test, the areas of tension noted by Douglas (2001b) (i.e. test authenticity and inseparability), as discussed in the literature review, were evident in the findings of this study. The thematic analysis revealed that some of the domain expert participants questioned the authenticity of the OET Writing task when
compared to their own workplace experience (i.e. some aspects of the task were not relevant to their individual current workplace practices). In addition, the qualitative analysis demonstrated that again some panellists at times struggled with focussing on language features alone as they were requested to do. These are issues that will likely remain of concern to LSP test developers. A further practical contribution of this study is that it has proposed more defensible standards for the OET Writing sub-test than those that are currently in place. As noted, the basis for the present OET standards has not been made public or open to scrutiny for some time. This study has also highlighted various issues with the OET Writing sub-test that point to avenues for improvement in the longer term. These are considered in the future research areas section below.

6.5 Limitations

This section presents the limitations of the study arranged around four themes: sampling of participants; sampling of scripts; comparative cut scores; and standard-setting procedures.

6.5.1 Sampling of participants

A limitation of the study is that the participants were all recruited from an inner-city, urban area (Melbourne, Australia) and from a somewhat narrow range of specializations/backgrounds. Consequently, it might prove challenging to claim that the panellists were representative of all domain expert doctors. While this lack of representativeness is likely to be true, to an extent, for any standard-setting panel, ideally there should be a mix of locations in a standard-setting study that includes urban, suburban and rural health professionals and with more varied participant backgrounds. To achieve this, it would be necessary to conduct multiple standard-setting workshops with many doctor panellists from around the region or countries served by the test in question. The practicability of this may be unfeasible for a single researcher due to logistical, time and financial constraints, but may be more manageable for commercial testing agencies.
6.5.2 Sampling of scripts

The choice of tasks themselves might have been a factor influencing the final cut score results as the tasks provided by Cambridge Boxhill Language Assessment were retired OET Writing sub-tests. It is debatable as to whether different results would have been obtained if different tests were used. Some judges commented that one of the task’s case notes (task two) seemed more complex/difficult than the other (task one), however Cambridge Boxhill Language Assessment states that all tasks are intended and designed to be of equal difficulty. Perceived differences in task complexity may nevertheless have had some impact for some standard-setting participants on their final performance level judgements. More evidence of statistical parallelism across tasks should ideally be made available to future researchers to ensure that judgements of readiness in particular are indeed transferrable to other versions.

6.5.3 Comparative cut scores

Another consideration of this study is that the OET is presently used for 12 health professions, however the focus of this research was only on doctors. At the moment, standards are common across professions and whether this is also justified in the case of the new standards is an open question. While the cut scores derived from this study have been supplemented by new nursing cut scores, proposed as part of the larger ARC study (i.e. Knoch et al., 2017), these may not be regarded as generalizable and representative of the views of all health professions. This can be resolved by undertaking replication standard-setting studies for each profession. This is now a matter for Cambridge Boxhill Language Assessment and has been recommended in the final project report (Knoch et al., 2017).

6.5.4 Standard setting procedures

An essential element of standard-setting procedures is for subject-matter experts to apply their own ‘indigenous assessment criteria’ when forming performance level judgements and not to be given any predetermined criteria. The subsequent
standards generated from the process should be formulated on participants’ own individual experience and knowledge from the work environment. However, some participants in this study stated that their preference would have been for more guidance or set benchmarks to establish performance judgements more successfully. Some participants felt that if this guidance had been given by the study’s moderator in the training sessions, they would have been more confident in setting standards. As the domain experts in this study were not language experts, some participants requested that pre-established examples of pass or fail samples could be given and discussed with panellists prior to training judgements being determined individually. However, this would require longer training sessions, further allocated resources and more importantly is not one of the purposes of standard setting – judgements should always be made without prior input from a convener as only domain experts themselves have the requisite workplace experience and knowledge to make appropriate performance level decisions. Thus, there is an implication that perhaps some participants did not fully understand the true nature of a standard-setting procedure in spite of what they professed when completing their evaluation forms. Further consideration needs to be given to ways of both explaining to domain experts why their ‘unfettered’ judgements are being sought and to ascertaining that they have understood the reasons for this.

The data in this study was drawn from seven different workshop sessions and five think-alouds, hence some variation was clearly evident (see qualitative examples in results section 4.2.11). It is conceivable that if a single workshop were conducted with participants the results may have been different. This is due to potential issues with the application and consistency of the chosen standard-setting method across sessions. However, steps were taken by the session facilitator to ensure that the method was applied as consistently as possible with all panel groups. In the interests of consistency, both within and across standard-setting studies, it would be desirable for researchers to refine their protocols on the basis of experience in conducting standard-setting sessions and to share these protocols with other researchers.
Another limitation, is related to the assumption “there is sufficient and appropriate information for the standard-setting panellists to make a decision” from Knoch and Macqueen’s (in preparation) argument-based validity framework for an LSP test. Some researchers of second language writing assessment (see for example, Lee, Gentile & Kantor 2008; Schoonen 2005) have found that more than one test candidate writing sample, in separate tasks, is necessary to establish a reliable and valid score that is representative of an individual’s overall writing capability. This is in contrast to the case of the OET Writing sub-test, as the OET only elicits a single sample of a test candidate’s writing. This is a potential limitation of the OET Writing sub-test itself and may be a consideration worth exploring through further test development of writing tasks.

6.6 Future research areas

From the lack of published standard-setting studies concerning English language proficiency in relation to language tests in general, and specific purpose tests, it seems apparent that more research is required. In addition, as Cizek (2012) mentions, the apparent randomness and lack of consistency in the application of standard-setting methods being employed is one of the broader problems. Furthermore, there may be replicability issues associated with any given cut score (Cizek, 2012) and a further consideration might be as to whether the use of another panel with the same method would have produced similar/different cut score results. Therefore, further replicable studies using a particular method (such as the AJM) and additional comparative methodological studies are warranted to assess whether this would have produced different or similar results in cut scores. Moreover, further standard-setting studies into health-specific tests such as the OET are necessary to make the system more transparent and accountable to the public and to shed further light on the standard-setting process. As noted, additional replication standard-setting studies for each profession on the OET should be undertaken. To further investigate whether domain experts are able to successfully determine valid standards on a language test (as non-language experts), future standard setting studies and procedures might consider comparative methods and employ language experts and subject-matter experts.
alongside one another, thereby facilitating the utilization of complementary expertise. In addition, further standard-setting studies with a qualitative focus are also needed. The use of think-aloud protocols (TAPs) was a unique aspect to this study that had been called for by previous standard-setting researchers. Their successful usage in this study indicates that the technique has much promise in better understanding participants’ thought processes while making standard-setting judgements. Hence, studies that utilize TAPs are necessary to make comparisons of their effectiveness. Further studies using TAPs would also help to better understand the construct underlying judgements and hence the validity of cut scores.

6.7 Summary

The study considered whether domain experts are best placed (as opposed to language experts) for setting the standards in a specific-purpose language (LSP) test such as the OET. The validity implications of these findings for the OET Writing sub-test, and for LSP testing more generally, were considered using a unique argument-based validity approach as proposed by Knoch and Macqueen (in preparation). Validity evidence gathered from this study confirmed, in the main, that the resulting new standards, as informed by domain expert involvement, were indeed warranted. By utilizing a relatively under-researched qualitative approach this study has made a further significant contribution to standard-setting research. Future standard-setting studies should also employ similar qualitative components to better understand participants’ thought processes and their relevance to underlying test constructs when making performance judgements.
References


Hunt, K. (1965). *Grammatical structures written at three grade levels*. NCTE research report No. 3. Champaign, IL: NCTE.


Taylor, C. (2007). Relating the Trinity College London International ESOL examinations to the CEFR. In *The Colloquium served very well the mission and aims of the Council of Europe.* (p. 69).


Appendices

Appendix A – Test prompt and writing sample, Task 1 ‘Brian Edwards’ (T1-BE)

**OCCUPATIONAL ENGLISH TEST**

**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:**

Mr Brian Edwards, born on 1st May, 1948, is a patient in your General Practice.

**Patient Details**

- **Name:** Mr Brian Edwards
- **Residence:** 42 Grand View Drive
  Mountain Valley
- **Social background:**
  85-year-old retired accountant
  Married, lives at home with wife
  Non-smoker, non-drinker

**Patient History**

**Date:** 12.06.2011

**Subjective:**

- **Presenting complaint:** 12/12 HS, progressively enlarging skin lesion on L. lower leg
  Associated 2/2 HS, swelling, 
  erythema
  Patient is anxious
  No pain/tenderness; no bone pain
  No LOA: no fever, no change in bowel habits
  Otherwise healthy, independent in all aspects of daily living

**Past medical Hx:**

- **2008:** Glaucoma
  Treatment changed

**Medications:**

- Tetracycline 0.25% solution; 1 drop twice daily
  No known allergies

**Family Hx:**

- No skin cancer in family

**Objective:**

- **T:** 98.7°C; P: 80; RR: 16; Ht: 175cm; Wt: 78kg
  Alert and orientated
  90% skin lesion on L. leg area
  Irregular edge, ulcerated, erythematous, purulent discharge
  Systemic review – Normal

**Assessment:**

- **2, 500cc**

**Plan:**

- Refer to plastic surgeon for assessment (excision) and excision
  Counsellor on surgery inc, own costs and provide reassurance that lesion is most probably localised
  Start antibiotics

**Writing Tasks:**

Using the information given in the case notes, write a letter or referral to the plastic surgeon, Mr. John Law, seeking follow-up assessment.

Address the letter to Mr. John Law, Department of Plastic Surgery, Main Hospital, Royal Avenue, Newtown.

**In your answer:**

- Expand the relevant notes into complete sentences
- Do not use note form
- Use letter format

The body of the letter should be approximately 100–200 words.
OCCUPATIONAL ENGLISH TEST

WRITING SUB-TEST: MEDICINE
SAMPLE RESPONSE: LETTER OF REFERRAL

Mr Jon Liew
Department of Plastic Surgery
Main Hospital
Royal Avenue
Newtown

12 March 2011

Re: Mr Brian Edwards
42 Grandview Drive
Mountain Valley
DOB: 16.05.46

Dear Mr Liew,

I am writing to refer Mr Edwards to you for follow-up assessment. Mr Edwards is a 65-year-old retired accountant who lives at home with his wife.

Mr Edwards presented today with a skin lesion on his left tibial area. Apparently it has progressively enlarged over 12 months. Two weeks ago, the lesion became red and swollen. On examination today, the lesion is of 9 cm diameter with an irregular edge, ulceration, erythema and purulent discharge.

Mr Edwards denies experiencing associated tenderness or bony pain, loss of weight, fever or change of bowel habits. A systems review was unremarkable. Vital signs are normal and Mr Edwards appears otherwise well.

Mr Edwards had a squamous cell carcinoma removed from his right tibial area in 2008 and I suspect that he has now developed another squamous cell carcinoma. This time the lesion appears to be infected.

Mr Edwards is feeling anxious about developing another skin cancer so I have reassured him that the lesion appears localized, with no signs of malignant spread. I have counselled him on surgery and the likelihood of him requiring a skin graft. I have started him on oral flucloxacillin.

Please don’t hesitate to contact me with any queries.

Yours sincerely,

Doctor
Appendix B – Test prompt, Task 2 ‘Betty Johnson’ (T2-BJ)

OCCUPATIONAL ENGLISH TEST

WRITING SUB-TEST: MEDICINE

TIME ALLOWED: 6 MINUTES

WRITING TIME: 40 MINUTES

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

Notes:

Your patient, an 81-year-old woman, recently had a right total knee replacement (R: TKR) on 25/02/2015. She is being discharged today.

Patient:
Ms Betty Johnson

Address:
12 Mary Street, Stillwater

Marital Status:
widowed

Past Medical History:
Aortic valve replacement & pacemaker 2010
Cataract surgery since 2011 — pain & immobility — past 2 yrs
For R:TKR Feb 2015: full blood work, typing & cross matching, X-rays, ECG etc.

Regular Medication (25/2/15):
Pantectol 35mg 2 tabs tabs
Warfarin 3mg once a day — crossed 5 days preoperatively, started Clexane (enoxaparin sodium — anticoagulant)

Social Background:
widowed 1985. Lives alone. 4 children

Past Op:
25/02/15 11:30am
Returned to ward following R:TKR.
Vital signs — BP 115/70, P 82, R 16, T 36.5°C.
Oxygen saturation good, lines inserted. i/v on pillow. Hb 8kg/dl — i/v Transfusion. IV cefuroxim 1g i/v 24 hours
Intravenous regular oral pantectol (1g edti)
Patient Controlled Analgesia (PCA) — morphine effective.
Wounds — ok.

26/02/15
Wound — good, sponged.
Restart warfarin 5mg today.
s/c Clexane 50mg given for anticoagulation.
Cefuroxim 1g — i/v 24 hours
Pathology: FBE, UA, Liver Function Tests (ALT, AST), Hb.
Path results: Hb 100g/dl — commence Fertab (iron sulphate) 300mg mane.

27/02/15
s/c Clexane 50mg.
Start warfarin 5mg nodes.
Removal of (RFC) dressing, wound good, RFC 3x clips on 03/03/15.

28/02/15
Catches, short walks. Wound good, satisfactory.
s/c Clexane 50mg given.

01/03/15
s/c Clexane 50mg given.

02/03/15
X-rays, bloods — INR — 3.0, Hb 111g/dl.
Pathology: no abnormalities.
Managing w/in assistance.
Cese Clexane.

03/03/15 — 05/03/15
Wound clean, RFC 3x clips tomorrow. Mobility good. obs ✓

06/03/15
RFC remaining clips. Pathology ✓ Transfer to rehab today.

Physio:
07/03/15
Admission complete — stable. Obs ✓ Mobility, catches good.

08/03/15 — 10/03/15
Mobility, frame use, trial stick, pool, gentle exercises ✓ good. Showering w/in assistance.
Path & X-ray.

14/03/15
Path ✓ INR — 3.8
Warfarin 4mg node. Hb — 112g/dl. Fertab 150mg mane.

15/03/15 — 16/03/15
Universal — gradually ✓ independence.
Wound good. Obs ✓ Physio exercises good. Home list provided.

21/03/15
No cardiac issues.
Discharged home nursing assistance (personal hygiene, home care). Wound expector, shower w/in assistance. Stick / frame.
Discharge medication: warfarin 4mg node, Fertab 150mg mane, pantectol 1g od, oxycodeone 5–10mg per.
Rehab setp in 2 weeks.
Advised to see local doctor in 1 week, referral for local doctor — suggest repeat FBE, INR.

Writing Task:

Using the information given in the case notes, write a letter of referral to Ms Johnson’s local doctor.

Address the letter to Dr Tony Jones, Private Practice, 12 New Street, Stillwater.

In your answer:

• Expand the relevant notes into complete sentences
• Use past tense form
• Use letter format

The body of the letter should be approximately 166–200 words.
Appendix C – AJM training response form

You will read a selection of five referral letters written by overseas-qualified health professionals (HPs). As if you were the supervisor of each HP, consider each writing sample in terms of their competence to participate in entry-level/ supervised clinical practice involving interaction with co-workers, supervisors and other HPs.

Please give an overall/holistic rating of the written communication skills in each response. Place an X in a category for each ‘Writing response ID code.’ If you cannot decide between levels, use the ‘between’ categories as freely as the others.

<table>
<thead>
<tr>
<th>Performance Level</th>
<th>Writing Response ID Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Please write the code of each response you rate here (e.g. T1-BE-92 or T2-BJ-17 from the top right of the response)</td>
<td></td>
</tr>
<tr>
<td>STRONG</td>
<td></td>
</tr>
<tr>
<td>✗ between STRONG and COMPETENT</td>
<td></td>
</tr>
<tr>
<td>✗ COMPETENT</td>
<td></td>
</tr>
<tr>
<td>✗ between COMPETENT and NOT YET COMPETENT</td>
<td></td>
</tr>
<tr>
<td>✗ NOT YET COMPETENT</td>
<td></td>
</tr>
<tr>
<td>✗ between NOT YET COMPETENT and UNSATISFACTORY</td>
<td></td>
</tr>
<tr>
<td>✗ UNSATISFACTORY</td>
<td></td>
</tr>
</tbody>
</table>
Appendix D – Final AJM take-home response form (seven categories)

You will read a selection of 30 referral letters written by overseas-qualified health professionals (HPs). As if you were the supervisor of each HP, consider each writing sample in terms of the HP’s competence to participate in entry-level/supervised clinical practice involving interaction with co-workers, supervisors and other HPs.

Please give an overall/holistic rating of the written communication skills in each response. Place an X in a category for each ‘Writing response ID code.’ If you cannot decide between levels, use the ‘between’ categories as freely as the others.

<table>
<thead>
<tr>
<th>Performance Level</th>
<th>Writing Response ID Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Please write the code of each response you rate here (e.g. T1-BE-92 or T2-BJ-17 from top right of response)</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>STRONG</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>between STRONG and COMPETENT</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>COMPETENT</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>between COMPETENT and NOT YET COMPETENT</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>NOT YET COMPETENT</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>between NOT YET COMPETENT and UNSATISFACTORY</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>UNSATISFACTORY</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
<tr>
<td>☐</td>
<td></td>
</tr>
</tbody>
</table>
Appendix E – Alternate AJM response form (12 categories)

You will read a sample of referral letters written by doctors who have a qualification from outside Australia. Consider this writing performance for the purpose of the doctor’s participation in **entry-level / supervised clinical practice** involving interaction with co-workers, supervisors and other health professionals, and as if you were his/her supervisor.

Please give an overall rating of the written communication skills in each response. Place an X in a suitable category for each ‘Writing response ID code’.

<table>
<thead>
<tr>
<th>Writing Response ID</th>
<th>UNSATISFACTORY</th>
<th>NOT YET COMPETENT</th>
<th>COMPETENT</th>
<th>STRONG</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
<td>Med</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix F – Performance Profile method response form

You will read a sample of referral letters written by doctors who have a qualification from outside Australia. Consider this writing performance for the purpose of the doctor’s participation in entry-level / supervised clinical practice involving interaction with co-workers, supervisors and other health professionals, and as if you were his/her supervisor.

The responses have been arranged, based on actual test takers real scores, from lowest to highest.

Study the responses carefully and working through them one by one, independently decide on a response as being minimally competent or just qualified candidate for the performance level and where you would ‘draw the line’ between performance levels.

Use the attached form to decide Yes (Y) or No (N) for each ‘Response ID code’ as being minimally competent or just-qualified.

<table>
<thead>
<tr>
<th>Actual test score (not known to participants)</th>
<th>Writing Response ID</th>
<th>Minimally Acceptable - Yes or No?</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.87</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>4.85</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>4.84</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>4.83</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>4.81</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>4.79</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>4.77</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>4.75</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4.73</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>4.72</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
Appendix G – ‘Think-Aloud’ pre-task form

‘Think-Aloud’ Pre-task Form

Thank you for agreeing to further participate in this study. To investigate what is important to you while you are judging OET test candidate writing responses, at the upcoming session I would like you to ‘talk out loud’ about what goes through your mind as you are forming judgements. There are no ‘right’ or ‘wrong’ judgements, but you will be asked to justify why you placed a response into a particular category.

As you read and judge each script, you will be asked to say all your thoughts ‘out loud’. At first this may seem odd, but will feel less strange as you become more familiar with the task. Your comments should be on the spur-of-the-moment and unprompted. Don’t be concerned if you feel hesitant, speak in incomplete sentences or change your mind about something.

You should try to include comments on what aspects capture your attention and the factors you take into consideration in deciding on a performance level. Try to give as much detail as you can and elaborate by pointing out examples and anything specific that contributes to your decisions. You should try to keep up a steady ‘stream of consciousness’ without stopping to think too much about what you would like to say.

Please view the ‘think-aloud’ example by clicking on the following link. It is an example of a doctor testing a new web page’s usability: https://www.youtube.com/watch?v=nJ2udLjdsx4
Appendix H – ‘Think-Aloud’ response form and guidelines

As in stage 1-2 of the study, you will read a selection of referral letters written by overseas qualified HPs whose first language is not English. As if you were the supervisor of each HP, consider each writing response in terms of the HP’s competence to participate in entry-level/supervised clinical practice involving interaction with co-workers, supervisors and other HPs.

Please give an overall/holistic rating of the written communication skills in each response. Place an X in a category for each ‘Writing response ID code.’ If you cannot decide between levels, use the ‘between’ categories as freely as the others.

‘Think Aloud’ Protocol

1. To investigate what is important to you while you are judging OET test candidate writing samples, I would like you to ‘talk out loud’ about what goes through your mind as you are forming judgements.

2. While you are reading each response, please say all your thoughts ‘out loud’. At first this may seem odd, but will feel less strange as you become more familiar with the task. Your comments should be on the spur-of-the-moment and unprompted. Don’t be concerned if you feel hesitant, speak in incomplete sentences or change your mind about something.

3. You should try to include comments on what aspects capture your attention and the factors you take into consideration in deciding on a performance level. Try to give as much detail as you can and elaborate by pointing out examples and anything specific that contributes to your decisions. You should try to keep up a steady ‘stream of consciousness’ without stopping to think too much about what you would like to say.

4. Before you begin reading each writing response, please read out the sample number in the top right-hand corner. You should address your comments to yourself (or the digital recorder).
<table>
<thead>
<tr>
<th>Level</th>
<th>Writing Response ID Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Please write the code of each response you rate here (e.g. T1- BE-92 or T2-BJ-17 from top right of the response)</td>
<td></td>
</tr>
<tr>
<td>STRONG</td>
<td></td>
</tr>
<tr>
<td>☧ between STRONG and COMPETENT</td>
<td></td>
</tr>
<tr>
<td>☩ COMPETENT</td>
<td></td>
</tr>
<tr>
<td>☩ between COMPETENT and NOT YET COMPETENT</td>
<td></td>
</tr>
<tr>
<td>☩ NOT YET COMPETENT</td>
<td></td>
</tr>
<tr>
<td>☩ between NOT YET COMPETENT and UNSATISFACTORY</td>
<td></td>
</tr>
<tr>
<td>☩ UNSATISFACTORY</td>
<td></td>
</tr>
</tbody>
</table>
### Appendix I – Workshop attendees by medical practitioner specialisation, current/recent workplace(s) and role(s), years of experience since registration in Australia, and years of experience in a supervisory role

<table>
<thead>
<tr>
<th>Medical practitioner specialisation</th>
<th>Current/recent workplace(s) and role(s)</th>
<th>Years of experience since registration in Australia</th>
<th>Years of experience in a supervisory role</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judge 1</td>
<td>Oncology</td>
<td>Director of cancer services at a public hospital &amp; private practice</td>
<td>39</td>
</tr>
<tr>
<td>Judge 2</td>
<td>General Practitioner (GP)</td>
<td>GP at a private medical centre</td>
<td>8</td>
</tr>
<tr>
<td>Judge 3</td>
<td>Oncology</td>
<td>Oncology research fellow &amp; registrar at a public hospital</td>
<td>7</td>
</tr>
<tr>
<td>Judge 4</td>
<td>Palliative Care</td>
<td>Director of palliative care at a public hospital</td>
<td>26</td>
</tr>
<tr>
<td>Judge 5</td>
<td>General Practitioner (GP)</td>
<td>GP at a private medical centre</td>
<td>4</td>
</tr>
<tr>
<td>Judge 6</td>
<td>General Practitioner (GP)</td>
<td>GP at a private medical centre</td>
<td>19</td>
</tr>
<tr>
<td>Judge 7</td>
<td>General Practitioner (GP)</td>
<td>GP at a private medical centre</td>
<td>6</td>
</tr>
<tr>
<td>Judge 8</td>
<td>Endocrinology</td>
<td>Senior endocrinology consultant at a public hospital</td>
<td>27</td>
</tr>
<tr>
<td>Judge 9</td>
<td>Palliative Care</td>
<td>Palliative care consultant at a public hospital</td>
<td>19</td>
</tr>
<tr>
<td>Judge 10</td>
<td>Dermatology &amp; Medical Education</td>
<td>Dermatologist at a skin cancer foundation &amp; medical education fellow at a public university</td>
<td>18</td>
</tr>
<tr>
<td>Judge 11</td>
<td>General Practitioner (GP)</td>
<td>GP at a public university health service</td>
<td>39</td>
</tr>
<tr>
<td>Judge 12</td>
<td>Palliative Care</td>
<td>Palliative care consultant at a public hospital</td>
<td>42</td>
</tr>
<tr>
<td>Judge 13</td>
<td>Palliative Care</td>
<td>Palliative care consultant at a public hospital</td>
<td>8</td>
</tr>
<tr>
<td>Judge 14</td>
<td>Medical Education</td>
<td>Sub-dean of medical education at a public university</td>
<td>35</td>
</tr>
<tr>
<td>Judge 15</td>
<td>General Practitioner (GP) &amp; Medical Education</td>
<td>GP in remote indigenous health &amp; senior lecturer in medical education at a public university</td>
<td>20</td>
</tr>
<tr>
<td>Judge 16</td>
<td>Paediatric Endocrinology</td>
<td>Paediatric endocrinologist at a public hospital &amp; private practice</td>
<td>16</td>
</tr>
<tr>
<td>Judge 17</td>
<td>General Practitioner (GP)</td>
<td>Medico-legal advisor at a public university health service</td>
<td>24</td>
</tr>
<tr>
<td>Judge 18</td>
<td>General Practitioner (GP)</td>
<td>GP at a private medical centre</td>
<td>18</td>
</tr>
</tbody>
</table>
### Appendix J – Workshop attendees by supervision of an overseas-trained medical practitioner, years/months of overseas-trained medical practitioner supervision, supervision of non-native English speaker/s

<table>
<thead>
<tr>
<th></th>
<th>Supervision of an overseas-trained medical practitioner – Yes/No</th>
<th>Years/months of overseas-trained medical practitioner supervision</th>
<th>Non-native English speaker supervision – Yes/No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judge 1</td>
<td>Y</td>
<td>5 years</td>
<td>Y</td>
</tr>
<tr>
<td>Judge 2</td>
<td>Y</td>
<td>1 year</td>
<td>Y</td>
</tr>
<tr>
<td>Judge 3</td>
<td>N</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Judge 4</td>
<td>Y</td>
<td>3 months</td>
<td>Y</td>
</tr>
<tr>
<td>Judge 5</td>
<td>N</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Judge 6</td>
<td>N</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Judge 7</td>
<td>N</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Judge 8</td>
<td>Y</td>
<td>undetermined</td>
<td>Y/N (both)</td>
</tr>
<tr>
<td>Judge 9</td>
<td>Y</td>
<td>15</td>
<td>Y/N (both)</td>
</tr>
<tr>
<td>Judge 10</td>
<td>N</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Judge 11</td>
<td>Y</td>
<td>6 months</td>
<td>Y</td>
</tr>
<tr>
<td>Judge 12</td>
<td>Y</td>
<td>6 months</td>
<td>Y</td>
</tr>
<tr>
<td>Judge 13</td>
<td>N</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Judge 14</td>
<td>N</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Judge 15</td>
<td>N</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Judge 16</td>
<td>Y</td>
<td>undetermined</td>
<td>Y/N (both)</td>
</tr>
<tr>
<td>Judge 17</td>
<td>N</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Judge 18</td>
<td>N</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### Appendix K – Workshop level of workplace interaction with non-native English-speaking colleagues

<table>
<thead>
<tr>
<th></th>
<th>Limited</th>
<th>Moderate</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judge 1</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Judge 2</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Judge 3</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Judge 4</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Judge 5</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Judge 6</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Judge 7</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Judge 8</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Judge 9</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Judge 10</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Judge 11</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Judge 12</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Judge 13</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Judge 14</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Judge 15</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Judge 16</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Judge 17</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Judge 18</td>
<td>X</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix L – Number of referral letters (or similar pieces of communication) written by a non-native English speaking medical practitioner read per month

<table>
<thead>
<tr>
<th>Judge</th>
<th>0-2</th>
<th>3-5</th>
<th>6-10</th>
<th>10+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judge 1</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Judge 2</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Judge 3</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Judge 4</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Judge 5</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Judge 6</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Judge 7</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Judge 8</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Judge 9</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Judge 10</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Judge 11</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Judge 12</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Judge 13</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Judge 14</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Judge 15</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Judge 16</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Judge 17</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Judge 18</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

Appendix M – Workshop attendees by country of birth, native language/s, country where majority of medical training was undertaken

<table>
<thead>
<tr>
<th>Judge</th>
<th>Country of birth</th>
<th>Native Language/s</th>
<th>Country where majority of medical training was undertaken</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judge 1</td>
<td>Israel</td>
<td>English</td>
<td>Australia</td>
</tr>
<tr>
<td>Judge 2</td>
<td>Australia</td>
<td>English</td>
<td>Australia</td>
</tr>
<tr>
<td>Judge 3</td>
<td>Hong Kong</td>
<td>Chinese (Cantonese)</td>
<td>Australia</td>
</tr>
<tr>
<td>Judge 4</td>
<td>Australia</td>
<td>English</td>
<td>Australia</td>
</tr>
<tr>
<td>Judge 5</td>
<td>Zimbabwe</td>
<td>Shona/English</td>
<td>Zimbabwe/UK</td>
</tr>
<tr>
<td>Judge 6</td>
<td>Hong Kong</td>
<td>Chinese (Cantonese)</td>
<td>Australia</td>
</tr>
<tr>
<td>Judge 7</td>
<td>China</td>
<td>Chinese (Mandarin)</td>
<td>Australia</td>
</tr>
<tr>
<td>Judge 8</td>
<td>Australia</td>
<td>English</td>
<td>Australia</td>
</tr>
<tr>
<td>Judge 9</td>
<td>Australia</td>
<td>English</td>
<td>Australia</td>
</tr>
<tr>
<td>Judge 10</td>
<td>Australia</td>
<td>English</td>
<td>Australia</td>
</tr>
<tr>
<td>Judge 11</td>
<td>Sri Lanka</td>
<td>English</td>
<td>Sri Lanka</td>
</tr>
<tr>
<td>Judge 12</td>
<td>Australia</td>
<td>English</td>
<td>Australia</td>
</tr>
<tr>
<td>Judge 13</td>
<td>Singapore</td>
<td>English</td>
<td>UK</td>
</tr>
<tr>
<td>Judge 14</td>
<td>Australia</td>
<td>English</td>
<td>Australia</td>
</tr>
<tr>
<td>Judge 15</td>
<td>Scotland</td>
<td>English</td>
<td>Australia</td>
</tr>
<tr>
<td>Judge 16</td>
<td>Australia</td>
<td>English</td>
<td>Australia</td>
</tr>
<tr>
<td>Judge 17</td>
<td>England</td>
<td>English</td>
<td>UK</td>
</tr>
<tr>
<td>Judge 18</td>
<td>South Africa</td>
<td>English</td>
<td>Australia</td>
</tr>
</tbody>
</table>
Appendix N – Plain Language Statement (PLS) and Consent form

Project title: Investigating and revising the standards set on the Occupational English Test’s (OET) writing sub-test

Participants are sought for the above interdisciplinary study being conducted at the University of Melbourne.

<table>
<thead>
<tr>
<th>Responsible researcher:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dr Ute Knoch</td>
</tr>
<tr>
<td>School of Languages and Linguistics</td>
</tr>
<tr>
<td>University of Melbourne</td>
</tr>
<tr>
<td>Parkville, VIC, 3010</td>
</tr>
<tr>
<td>email: <a href="mailto:uknoch@unimelb.edu.au">uknoch@unimelb.edu.au</a></td>
</tr>
<tr>
<td>Telephone: (03) 8344 5206</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Other researchers involved in project:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assoc. Prof. Cathie Elder</td>
</tr>
<tr>
<td>Assoc. Prof. Robyn Woodward-Kron</td>
</tr>
<tr>
<td>Prof. Tim McNamara</td>
</tr>
<tr>
<td>Prof. Elizabeth Manias</td>
</tr>
<tr>
<td>Assoc. Prof. Eleanor Flynn</td>
</tr>
<tr>
<td>Mr Simon Davidson (PhD candidate)</td>
</tr>
<tr>
<td>Ms Annemiek Huisman</td>
</tr>
</tbody>
</table>

What is the project about?

Overseas trained health professionals (HPs) who plan to practice in Australia need to show acceptable English language proficiency as part of the professional registration process. Concerns have been raised that the level of written English required on the language tests (including the Occupational English Test (OET), a specific-purpose language test for HPs), may be insufficient for effective functioning in the workplace. In response, the study will consider the minimum standards of performance on the writing sub-test of the OET. This is via the process of ‘standard setting,’ a systematic means of deciding what level of writing proficiency is deemed acceptable for professional practice by subject-matter experts (HPs). The overall purpose of this project is to revise the current standards of the OET writing sub-test and to investigate the thought processes of participants while setting these revised standards. The research is funded by the Australian Research Council and the OET Centre. This project has been approved by the University of Melbourne, Faculty of Arts, Human Ethics Advisory Group (HEAG) Project no: 1545516.1.

What will I be asked to do?

You will be asked to give judgements (using a designated check box form) about performance levels for writing test responses (180 – 200-word referral letters) from the OET. These were written by HPs who have a qualification from outside Australia and whose first language is not English. There will be 3 stages in the standard setting process. Stage 1 involves a workshop training session at a convenient time and location (60 mins). In stage 2 you will be asked to take home and judge 30 test candidate responses using a designated check box form (approx. 60 mins). Total time for stages 1 – 2 = 2 hours. You will be given a $400 gift card as a token of appreciation for your participation. In stage 3, as part of a further under-researched, qualitative section of the study, interested participants (optional only) will be asked to judge an additional 10 OET writing responses and record their thought processes while rating these scripts (approx. a further 45 – 60 mins). This will be at a convenient time and location for you. Total time for stages 1 – 3 = 2.75 – 3 hours. Interested participants who judge an extra 10 OET responses will be given an additional $100 gift card.
What happens next if I agree to participate?

If you agree to participate in the study, please contact Simon Davidson (PhD candidate) to register your interest/availability at s.davidson3@student.unimelb.edu.au or 0431 569 599. We will organize a convenient time and location for the training workshop. This will be arranged for you and other participants either at the University of Melbourne or an onsite training room. You will be sent a sample OET writing task and test candidate response to read before the workshop.

How will my confidentiality be protected?

The information you provide is confidential. Your anonymity and the confidentiality of your responses will be protected to the fullest possible extent within the limits of the law. Your name will not be used in any publications arising from the research and only the researchers will have access to the recordings from the sessions. Any references to personal information that might allow someone to guess your identity will be removed. The recording will be kept securely for 5 years and then destroyed.

What happens if I decide later that I don’t want to be involved?

Participation is voluntary. You can decide not to be involved at any time. In that case, we will not use the information you have given us and the data will be destroyed.

Where can I get further information?

Should you require any further information, or have any concerns, please do not hesitate to contact Dr Ute Knoch (details above). Should you have any concerns about the conduct of the project, you are welcome to contact the Executive Officer, Human Research Ethics, The University of Melbourne, on ph: 03 8344 2073, or fax: 03 9347 6739.
Consent Form
Health/Social Science Research – Adult providing own consent

Title
Investigating and revising the standards set on the 
Occupational English Test’s (OET) writing sub-test

Project Sponsor
Australian Research Council
Occupational English Test (OET) Centre

Coordinating Principal Investigator
Dr Ute Knoch (CPI)

Principal Investigators
A/Prof Catherine Elder
A/Prof Robyn Woodward-Kron
Prof Elizabeth Manias
A/Prof Eleanor Flynn
Prof Tim McNamara

Associate Investigators
Mr Simon Davidson (PhD candidate)
Ms Annemiek Huisman

Location
The University of Melbourne

Declaration by Participant
I have read the Plain Language Statement. I understand the purposes, procedures and risks of the 
research described in the project. I have had an opportunity to ask questions and I am satisfied with 
the answers I have received. I freely agree to participate in this research project as described and 
understand that I am free to withdraw at any time during the project without any consequences. I agree 
to keep the research tasks/writing samples confidential and delete/dispose of them securely. I 
understand the training workshop discussion will be recorded, transcribed, de-identified, kept securely 
for 5 years and then destroyed. I understand the information I provide will be treated confidentially 
and protected to the fullest possible extent, within the limits of the law. I understand that once signed 
this document will be retained by the researcher.

Name of Participant please print) __________________________________________
Signature ___________________________ Date __________________________

Declaration by Researcher†
I have given a verbal explanation of the research project, its procedures and risks and I believe that the 
participant has understood that explanation.

Name of Researcher† (please print) __________________________________________
Signature ___________________________ Date __________________________

† An appropriately qualified member of the research team must provide the explanation of, and 
information concerning, the research project.
Appendix O – Participant pre-task form

Project title: Investigating and revising the standards set on the Occupational English Test’s (OET) writing sub-test

Participant Pre-Task

Thank you for agreeing to participate in this study. During the training workshop you will be asked to give judgements about a selection of writing responses from the Occupational English Test’s (OET) writing sub-test. These were written by health professionals (HPs) who have a qualification from outside Australia and whose first language is not English. As if you were the supervisor of each HP, you will be asked to consider each writing sample in terms of the HP’s competence to participate in entry-level/ supervised clinical practice involving interaction with co-workers, supervisors and other HPs. You will be asked to judge candidate responses at a number of levels. There are no ‘right’ or ‘wrong’ judgements, but you may be asked to justify why you placed a response into a particular category. These are the descriptions that will be used for each performance standard level:

STRONG	COMPETENT	NOT YET COMPETENT	UNSATISFACTORY

Before you attend the workshop, please think about and come prepared to discuss what it means for an overseas-trained HP to have ‘minimally competent’ written English communication skills in the workplace i.e. in your view, what should a HP be able to demonstrate or ‘CAN DO’ in their written communication. What additional aspects would make them STRONG? What may they struggle with or couldn’t demonstrate if they were NOT YET COMPETENT? What other aspects would make their workplace written communication UNSATISFACTORY? You may bring some notes to the workshop if you wish.

STRONG

COMPETENT

NOT YET COMPETENT

UNSATISFACTORY
Appendix P – Participant background information form

Project title: Investigating and revising the standards set on the Occupational English Test’s (OET) writing sub-test

Participant Background Information

Name:        Gender:
Country of birth:     Native language:
Country where majority of medical training was undertaken:
Profession:      Specialisation:
Current/ recent workplace(s) and role(s):

Years of experience since registration in Australia:
If applicable, years of experience in a supervisory role:
Have you ever supervised an overseas-trained medical practitioner?       Yes/No
If so, how many years/months?
If so, were they native English speaker/s?     Yes/No
Level of workplace interaction with non-native English-speaking colleagues (please circle):
Limited    Moderate    High
On average, approximately how many referral letters (or similar pieces of communication) written by a non-native English speaking medical practitioner would you read per month (please circle):

0-2       3-5       6-10       10+

Your email address:
Preferred method for returning forms (please circle):
Via email (soft copy only)         Via email (hard copy scanned)         Via post
Your postal address for gift card delivery (if necessary):
Appendix Q – Training evaluation form

Training Evaluation

Name:

1. Please indicate the degree to which you agree with each of the following statements.

<table>
<thead>
<tr>
<th>Please place an ‘X’ in each box</th>
<th>Strongly Agree</th>
<th>Somewhat Agree</th>
<th>Neutral</th>
<th>Somewhat Disagree</th>
<th>Strongly Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) I understand the standard-setting task</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b) The training in the method was adequate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c) I understand the difference between performance levels including the ‘between’ levels</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d) I understand how to make judgements</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>e) I understand how to use the judgement form provided</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2. Please indicate the degree to which you agree with each of the following statements.

<table>
<thead>
<tr>
<th>Please place an ‘X’ in each box</th>
<th>Strongly Agree</th>
<th>Somewhat Agree</th>
<th>Neutral</th>
<th>Somewhat Disagree</th>
<th>Strongly Disagree</th>
<th>Not Applicable</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) The feedback and discussion helped to refine my understanding of the standard-setting process</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b) The feedback and discussion helped to refine my judgements and understanding of the performance levels</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c) The feedback and discussion will influence my further writing response judgements</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Appendix R – Final evaluation form

## Final Evaluation

Name:

1. Overall, how influential was each of the following in making your judgements? Please rate each statement using the given scale. Please place an ‘X’ in each box.

<table>
<thead>
<tr>
<th></th>
<th>Very Influential</th>
<th>Somewhat Influential</th>
<th>Not at all Influential</th>
<th>Not Applicable</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) My experience with the type of writing task from my working life</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b) My sense of what test candidates should be able to do in order to be proficient/ safe in the workplace</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c) My own definitions of performance levels</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d) Other participants’ definitions of performance levels</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>e) Feedback/comments/discussion with other participants</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>f) Writing response judgements/ratings of other participants</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2. How useful was each of the following in judging writing responses? Please place an ‘X’ in each box.

<table>
<thead>
<tr>
<th></th>
<th>Very Useful</th>
<th>Somewhat Useful</th>
<th>Not at all Useful</th>
<th>Not Applicable</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) Practicing the procedure</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b) Feedback/comments/discussion with other participants</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c) Writing response judgements/ratings of other participants</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3. Please indicate the degree to which you agree with each of the following statements. Please place an ‘X’ in each box.

<table>
<thead>
<tr>
<th></th>
<th>Strongly Agree</th>
<th>Somewhat Agree</th>
<th>Neutral</th>
<th>Somewhat Disagree</th>
<th>Strongly Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) I believe that from this study, I have a greater understanding of the importance and value of standard setting</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b) The standard-setting procedure was a worthwhile exercise for me</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c) I believe that in future, standard-setting sessions should be conducted at regular intervals</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4. Do you have any additional comments or feedback about the standard-setting process?

Thank you for your time and participation in the study.
Appendix S – Think-aloud final evaluation form

Think-Aloud Final Evaluation

Name:

1. Please indicate the degree to which you agree with each of the following statements. Please place an ‘X’ in each box.

<table>
<thead>
<tr>
<th></th>
<th>Strongly Agree</th>
<th>Somewhat Agree</th>
<th>Neutral</th>
<th>Somewhat Disagree</th>
<th>Strongly Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) The instructions were clear and I understood what I was</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>required to do when ‘thinking aloud’</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>b) The level of prompting from the researcher was adequate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c) I was comfortable with the number of writing responses I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>judged and spoke about</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d) I had adequate time to speak comprehensively about each</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>writing response</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>e) The process of ‘thinking aloud’ didn’t alter the way in</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>which I judged the writing responses</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2. Do you have any additional comments about the experience of ‘thinking aloud’ overall?

Thank you for your time and participation in the study.
Appendix T – Task 1 ‘Brian Edwards’, T1-BE-74 test candidate sample

Answer

Mr Jon Liew
Department of Plastic Surgery
Main Hospital
Royal Avenue
Newtown

12 March 2011

Dear Mr Liew

Thank you for seeing Mr Brian Edwards, a 65-year-old retired accountant who I suspected as suffering from Squamous cell cancer (SCC) on left lower leg

Mr Edwards presented today with a 12-month history of progressively enlarging skin lesion on left lower leg. He also noticed the skin lesion has become swelling and erythematous for the last 2 weeks. He is anxious because he had SCC lesion excised in 2008. He has no history of CVD or Diabetes and is married and lives with wife. He is no smoker neither a drinker. On examination, there was a 3 cm skin lesion on left tibia area associated with a rougher edge ulcerated, erythematous and purulent discharge. His vital signs and System review were normal. I suspected he has a SCC or possible an infection and commanded him on oral Flucloracillin

I would appreciated it if you could access him and provide his further management.

Yours Sincerely
Dr Q Chen
### Appendix U – Data dictionary

<table>
<thead>
<tr>
<th>Theme/Code</th>
<th>Category Description</th>
<th>Example</th>
<th>Code overlap in an utterance (if applicable – in bold)</th>
<th>Also relates to code (if applicable)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PL</td>
<td>Performance level judgement</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PL</td>
<td>Overall performance level judgement (summative)</td>
<td>I think this is “Competent” and probably “between Strong and Competent”. I put it “between Not Yet Competent and Unsatisfactory”.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TF</td>
<td>Task Fulfillment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td>Overall quality of the letter (summative)</td>
<td>Now I think this is quite a good letter. This is probably one of the better written letters.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>Overall meaning, clarity, intelligibility conveyed / or not i.e. meaning has still been conveyed even if grammar/ language errors are present or meaning is confused/ambiguous because of errors</td>
<td>It gets the meaning across very well. <strong>A few things with language that you have to look at again, but the meaning’s still reasonably clear.</strong></td>
<td>LG</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>Content</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PURP</td>
<td>Overall purpose of the letter evident/ or not</td>
<td>Why the referral’s being made and the purpose of it very clearly up front. I was gonna say the same thing about purpose of referral. That means that the... nice and clear, as long as that is seen in the letter I think that’s good communication.</td>
<td>PR</td>
<td>M</td>
</tr>
</tbody>
</table>
| UR | Urgency of the case presented evident/ or not | From the beginning for the person who is triaging to know if this cancer is urgent or just a worry about infection or something that has been there for a long time and probably has just got a little worse.

So your main concern and what you think it is in terms of urgency. |  |
|---|---|---|---|
| RC | Relevant content and ideas from patient case notes included i.e. general comments related to case note information being included | He’s got all the relevant details.

Well I guess I consider it “Competent” because it does convey all the information you need to know. | PL |
| IC | Irrelevant content and ideas from patient case notes included i.e. general comments related to case note information being included that should not have been, redundancy and repetition of ideas | Things like the heart rate aren’t, probably not, completely relevant.

I guess it’s partly depending on whether they feel they have to include all of this or whether they can leave some of it out, because some of it isn’t... the surgeons won’t care what the pulse rate is, really, or what the height and weight. But they’ve put that in. But, you know, it’s not wrong to put it, but, again, it’s maybe hiding the more relevant facts, just bogging it down with information |  |
| CA   | Content accuracy and transcription of case note content e.g. names, dates, addresses, ages, medications, diseases, symptoms etc. | He said he was ‘42’ which struck me and... WHAT WAS THE PROBLEM THERE? He got the age wrong. It’s actually his address number, ‘42’.

There are a few minor sort of spelling errors including the patient’s name which I just think is only a minor thing, but the patient’s name is pretty important. | SP |
| IM   | Information missing from case notes | However, there is no pain, fever or loss of weight or appetite. He has a personal history of squamous cell carcinoma of the pre-tibial skin. So, it doesn’t identify which side, but that’s a little bit annoying.

Again, I don’t know all the information. Where is it? I’m just looking. | |
| CE   | Clinical examination evident/ or not | Physical examination paragraph is all reasonably clear.

So fairly succinct examination findings compared to some of the others. | M CON |
| D    | Diagnosis/provisional diagnosis evident/ or not | It would be handy, I think, to put in, they put the provisional | PR |
diagnosis **at the end** and said: *My provisional diagnosis is a squamous cell carcinoma.*

They've said the provisional diagnosis **up front.**

**TP**  
Possible treatment plan evident/ or not

He’s telling the surgeon what he thinks should be done which is reasonable. He explained that he thought the surgeon would most likely remove it.

They’ve started them on Fluclox.

**CM**  
Current/ new medication listed/ or not e.g. Flucloxacillin (Fluclox)

Clear and correct information there and have the current medication, so that’s fine.

*I probably don’t feel I can quite say “Competent”, because I think that is important, not having the current medications.*

**RM**  
Regular medication listed/ or not

Goes on to mention regular medicines. That’s fine.

I think if they went to treat this person they need to know what medications they are on in terms of anaesthetic or risk of bleeding and stuff like that.

**PC**  
Presenting complaint evident/ or not i.e. relevant current/ past details of the presenting complaint being referred

But then they go into that, looks like, in the next paragraph: *Came to see me today*
complaining of a skin lesion on the lower leg which he's had over the last one year; the skin lesion has been progressively enlarging.

It's gone on to the history of the presenting complaint.

| MH | Relevant past medical history evident/ or not i.e. diseases/ symptoms/ conditions/ allergies to the presenting complaint | They don't have a separate paragraph for the past medical history and in terms of the glaucoma and the 'Timolol' and the absence of allergies. They've also drawn attention to the absence of important systemic features in the patient’s history. | DS/P |
| SH | Relevant past social/ family history evident/ or not e.g. if the patient is a smoker or not, their alcohol consumption, marital status, employment status etc. | He’s given a bit of a social history. They don’t mention stuff like not smoking that I can see which is relevant because it impacts on healing and things like that. | PR |
| PID | Patient identification mentioned generally as evident/ or not Patient identified by name = ID1 Patient identified by age or date of birth = ID2 Patient identified by address = ID3 | You know the demographic data of the patient, you know just really fundamental stuff, leaving them out, that would be, I would put that as “Unsatisfactory.” This one, common problem of lacking the date of birth, or address, or other identifying details | LO |
| DID | Doctor identification mentioned generally as evident/ or not e.g. writer/sender’s title, address, contact details, department etc. | Who is writing the letter, their job title, their address, their contact telephone number and possibly even an email address.  
They go to great lengths to write in his name and address [referring doctor] which is important for when a referral letter goes back, when the letter goes back, the name and address of the doctor he’s referring to which is probably not that important. | LO |
| RA | Receiver’s address included/ or not and where this is situated on the page | It’s a letter that’s dated and addressed to the surgeon.  
This is addressed to the surgeon, surgeon’s address | LO |
| E | Expression | It’s concise, covers what needs to be covered.  
Once again, if somebody is reading it quickly, it draws their attention to what’s already been done. | PR |
| CON | Conciseness, succinctness, efficiency of the letter for the reader evident/ or not | And his grammar is not that good, line 23, it should be ‘with’, ‘associated with.’  
So, they need to have not perfect grammar, but good enough so that you M | G | Accuracy of grammar, syntax and sentence structure |
| LG  | Language – general  
i.e. general summative comments about language use | So, the English is quite good.  
And I think the English reads pretty well. |
|-----|-----------------------------------------------------------------|----------------------------------------------------------------------------------|
| V   | Suitable use of vocabulary, appropriateness of word choice and wording evident/ or not | Again, a little bit of strange use of English on line 27,  
*that he doesn’t smoke, neither drinks alcohol.*  
The only other thing is just a few there’s a few English language things like,  
*I would be greatly appreciated for your assessment and further management,*  
rather than, ‘I would greatly appreciate...’ |
| MT  | Suitable use of appropriate medical terminology evident/ or not | They should be aware of words like heamoty, haemtosy or haemoterial.  
Technical terminology which is I think pretty important in medicine.  
Not all the correct language used, like ‘pussy’ you don’t say ‘pussy’, you say ‘purulent’, but that’s sort of terminology... |
| AB  | Suitable use of abbreviations evident/ or not | So, there may be times when they need to put in abbreviations, but then they need to know what the standard |

M
<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>abbreviations are that everyone knows and what’s appropriate and what’s not appropriate to put in any particular types of letter because I think there are some abbreviations that I would like to see and don’t want to see the whole word spelt out repeatedly.</td>
<td>I think inevitably in medical letters and communication abbreviations will crop up and I think as a general rule that should be kept at a minimum in letters and communication.</td>
<td></td>
</tr>
<tr>
<td>PT</td>
<td>Appropriate use of professional tone and register evident/ or not</td>
<td>I mean the information is communicated, they’re just not the proper way that you would be communicating with a colleague in written form, I think. There’s an informality about that. I don’t like saying the best assessment management from you. It’s a little bit sucking up to the person when you don’t need to. I think that tone, I don’t know if it’s a judgement thing, we just need to do our job.</td>
</tr>
<tr>
<td>O</td>
<td>Organisation</td>
<td></td>
</tr>
<tr>
<td>DS</td>
<td>Discourse structure of the letter</td>
<td>It’s competent in a sense that the</td>
</tr>
<tr>
<td></td>
<td>i.e. general comments related to the organization/sequence of ideas, stages, tasks, actions and processes</td>
<td>information’s been put down on the page, but in terms of what you’d expect of someone writing a referral letter, even at a fairly junior state I would expect it to be a bit better in terms of structure and that sort of thing. So that in reading it you can you know, see in a standard way, you know, that it’s conveyed in the correct sort of order and it’s not all over the place.</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>PR</td>
<td>Prioritizing key details/information for the reader i.e. putting more important information before less important information for the reader</td>
<td>You don’t want to have to find any importance in the second last sentence, hidden in a way that you miss it. It really has to say everything up front. Even if I was going to write it like this, I’d still put, I’d put the SCC first and then say, ‘Other than that they have glaucoma.’</td>
</tr>
<tr>
<td>P</td>
<td>Paragraphing of ideas evident/ or not i.e. general comments related to the use or not of paragraphs</td>
<td>I think it’s the set up because there’s a lot of, it’s like paragraphed. There’s no paragraphs here.</td>
</tr>
<tr>
<td>LOG</td>
<td>Use of logic evident/ or not</td>
<td>I mean in theory many of the letters that we see from local health professionals and many of the people whose first language is not</td>
</tr>
<tr>
<td>English may not have been trained in a system that respected that logic.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I’d expect those whose English wasn’t strong to still have thought about it logically.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| F | Overall flow of the letter i.e. general comments related to the flow of the letter | The actual flow of it is important.  
The flow is just, just sort of feels a bit disjointed. |
| PRES | Presentation | |
| LO | Overall appropriateness/ or not of letter layout/ format i.e. general comments about the use of standard formal letter layout | Just with the layout. Writing above the line, squeezing in words.  
The layout doesn’t help. If it was a little bit better laid out... |
| PRE | Overall appropriateness of letter presentation i.e. general comments about the overall presentation of the letter | I found it really messy. It doesn't look like a professional referral letter.  
I think there are some courtesy issues there for example, you don’t want a piece of scrappy paper, you don’t want dirty paper. It may not be the referrer’s fault but you certainly want something which looks clean and professional |
<p>| TS | Overall appropriateness/ or not of text spacing e.g. writing on every/ second line | I liked the fact that it had alternate lines so I could read it. |</p>
<table>
<thead>
<tr>
<th>CO</th>
<th>Crossed out information</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>i.e. general comments about words, sentences, paragraphs, sections being crossed out</td>
</tr>
<tr>
<td></td>
<td>When I initially look at this I notice there is a crossed-out paragraph which is a little bit annoying. So, it is a lot to read through and yeah with all the crossing out it took a while.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HW</th>
<th>Handwriting quality/readability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Handwriting is a little bit hard to read, but I can read it. Handwriting's a bit untidy.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PU</th>
<th>Punctuation and other mechanics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>i.e. general comments related to punctuation accuracy</td>
</tr>
<tr>
<td></td>
<td>He mixes up his capitals. Then, they've said, <em>He's come Today</em>... for some reason the ‘T’ is capitalised.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SP</th>
<th>Spelling</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>i.e. general comments related to spelling accuracy</td>
</tr>
<tr>
<td></td>
<td>Mr Edwards presented to me today with a ‘compaint’... so a spelling mistake. You don’t need to know how to exactly spell things.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PRO</th>
<th>Professionalism</th>
</tr>
</thead>
<tbody>
<tr>
<td>PK</td>
<td>Knowledge of professional role</td>
</tr>
<tr>
<td></td>
<td>i.e. information in the letter that demonstrates evidence of what a HP is/ isn’t able to do</td>
</tr>
<tr>
<td></td>
<td>He commenced <em>him on Timolol</em>. That’s unusual for a GP to ‘commence’ someone on Timolol. That would’ve been commenced by the eye specialist.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CC</th>
<th>Clinical competency and skills evident / or not</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Examination showed no signs of CE</td>
</tr>
<tr>
<td><strong>AA</strong></td>
<td>Audience awareness evident/or not i.e. an awareness of who the letter is intended for and how it might be read/viewed/interpreted by the reader</td>
</tr>
<tr>
<td><strong>PA</strong></td>
<td>Patient awareness evident/or not i.e. an awareness of/sensitivity to patient's situation, concerns etc.</td>
</tr>
<tr>
<td><strong>OTH</strong></td>
<td>Other</td>
</tr>
<tr>
<td><strong>EF</strong></td>
<td>Effort/Efficiency for the reader in dealing with the situation of</td>
</tr>
<tr>
<td>HO</td>
<td>A sense the letter is justified and not just a superficial handoff to the reader</td>
</tr>
<tr>
<td>TCB</td>
<td>Test candidate background e.g. local vs foreign, experienced vs novice, English as a first language vs ESL</td>
</tr>
</tbody>
</table>
| TC | General OET task comments | You’ve given them a 200-word limit. So, they won’t be writing too much. 
J15 But then it’s interesting because they were just looking at – they were just taking the notes, putting them into prose... 
J2 Taking – yeah. Exactly. It’s like | CON |
taking our consult notes and writing a letter from it.

J15
Which is not something any of us really do anymore. We just include those notes, like, with a summary opinion probably. Why can’t they use note format?
Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Davidson, Simon

Title:
How valid are domain experts' judgements of workplace communication? Implications for setting standards on the Occupational English Test (OET) Writing sub-test

Date:
2018

Persistent Link:
http://hdl.handle.net/11343/213877

File Description:
Full version- restricted access

Terms and Conditions:
Terms and Conditions: Copyright in works deposited in Minerva Access is retained by the copyright owner. The work may not be altered without permission from the copyright owner. Readers may only download, print and save electronic copies of whole works for their own personal non-commercial use. Any use that exceeds these limits requires permission from the copyright owner. Attribution is essential when quoting or paraphrasing from these works.