Comparing writing proficiency assessments used in professional medical registration: a methodology to inform policy and practice

Sathena Chan*, Lynda Taylor

*Centre for Research in English Language Learning and Assessment, University of Bedfordshire,
Hitchin Road, Luton, LU2 8LE, United Kingdom*

*Corresponding author, Tel.: +44 01582 489795
E-mail address: sathena.chan@beds.ac.uk

1. Introduction

Most health-care regulatory bodies in English speaking countries require overseas doctors to provide evidence of their English language proficiency, including writing skills, as part of formal registration requirements alongside proof of their medical expertise and skills. In the case of the UK, the General Medical Council (GMC) requires all doctors, including those who obtained their primary medical qualification within or outside the European Economic Area (EEA) (other than those who are exempt under certain legal conditions), and who are seeking professional registration in the UK, to demonstrate the necessary knowledge of English to ensure safe practice before registration and a licence to practise are granted. Unless they were able to provide alternative evidence, for many years, these international medical graduates (IMG), were required to take an English language proficiency (ELP) test - the Academic version of International English Language Testing System (IELTS), and to achieve a minimum overall score with minimum scores on the four subtests (i.e. Listening, Reading, Writing and Speaking)¹.

At the time the research reported here took place, IELTS was the only test that the GMC accepted as evidence of English language proficiency. However, other ELP tests, including both well-established products and more recent arrivals in the marketplace, were already accepted and used by other professional medical registration bodies worldwide where English is a key language. For the GMC this raised the ethical and practical question of whether one or more of these other ELP tests might be considered as acceptable alternatives alongside IELTS as satisfactory evidence of medical doctors' language proficiency in the UK. In light of this, the GMC issued a public tender for a research project to investigate this question so as to inform their future policy and practice. A comparative study was conducted by CRELLA and a full report made public in October 2015. A key outcome of the study was that, in 2018, the GMC opted to accept the Occupational English Test (OET) as capable of providing evidence of a doctor's language competency alongside IELTS (for more information, see <https://www.gmc-uk.org/news/news-archive/gmc-to-accept-new-english-language-qualification-for-non-uk-doctors>). As well as presenting the methodology used in the study, this paper reviews some current institutional policy and practice in this area to offer an up-to-date overview regarding the use of writing assessments in professional medical registration.

2. Literature review

2.1 Policy and practice on the regulation of overseas doctors

Studies on regulation policy and practice in the medical profession are relatively scarce. One such study conducted by RAND (<https://www.rand.org/>) in 2009 compared the systems in 10 regulatory medical authorities worldwide (i.e. Egypt, Germany, Greece, India, Italy, Nigeria, Pakistan, Poland, South Africa and Spain). Using a questionnaire, key informant interviews and comparative analysis, the study obtained useful insights on policy and practice for regulating overseas doctors in these countries. Here we highlight the major findings relevant to the discussion of this article. While medical regulatory systems can vary in significant ways between countries, registration with a medical regulatory body before doctors start practising is a common formal requirement. However, a formal revalidation system² for ensuring ongoing competence, including language competence, after qualification is uncommon in these countries. RAND (2009) cautioned that 'the costs of revalidation may be high, but it could be argued that revalidation will prevent certain (costly) fitness to practise

¹ The original GMC decision to adopt IELTS and identification of appropriate score levels is beyond the scope of this article (see Berry, O'Sullivan and Rugea, 2013).

² A quality assurance process in which registrants of a provincial medical regulatory authority are required to provide satisfactory evidence of their commitment to continued competence in their practice.

procedures and ensure patient safety' (p.26). Although the RAND study revealed useful information about the overall policy and practice of registration for overseas doctors, specific issues regarding doctors' proof of language for the exercise of their profession remain to be addressed. For example, what tests or evidence, when required, are used as proof of language proficiency in these countries? How do these tests compare with one another in terms of language demand and other test validity considerations? The present paper begins to try and answer these questions.

2.2 Issues in using ELP tests for medical registration purposes

Most previous studies in language assessment for the medical professions are concerned with: (i) the appropriateness (or authenticity) of a particular test (especially the speaking/writing components) (e.g. Elder, McNamara, Woodward-Kron, Manias, McColl, Webb and Pill, 2013; Knoch, McNamara, Woodward-Kron, Elder, Manias, Flynn and Zhang, 2015; McNamara, Elder, Flynn, Knoch, Manias, Woodward-Kron and Yahalom, 2018; Pill, 2016); and (ii) setting an appropriate cut score for proof of minimum required language proficiency (e.g. Berry, O'Sullivan and Rugea, 2013; Pill and McNamara, 2016). Some of these studies are outlined in more detail below.

The importance of matching the test to the target language use (TLU) domain is understandably regarded as a key consideration for test selection and use. Wette (2011) argued that most English proficiency tests are largely insufficient to assess the candidates' professional communication skills required in the TLU domain. To bridge the gap, she urged a collaborative approach involving specialists in English for Specific Purposes (ESP) and medical communication to identify candidates' developmental needs and provide appropriate courses of instruction. Pill's (2016) study to introduce workplace-specific communication criteria into OET is a good example of this approach. Through interviews with practitioners and thematic analysis, Pill investigated the construct of speaking skills required for healthcare consultations between trainee health professionals and patients in the TLU domain. Two professionally relevant criteria clinician engagement and management of interaction emerged from his findings. Other studies (see Knoch, et al., 2015; Knoch, Elder, Flynn, Manias, McNamara & Woodward-Kron, 2018; McNamara et al., 2018) addressed the OET writing component in a similar way through investigating the writing construct involved in generating patient records; identifying criteria indigenous to the work setting, applying these indigenous criteria to candidate performances on the OET writing task; and developing new passing standards against these criteria.

A second major consideration for medical regulatory bodies is the setting of appropriate cut scores for internationally trained medical graduates to join the profession and practise in a new jurisdiction. Berry, O'Sullivan and Rugea (2013) investigated the extent to which existing IELTS cut scores used by the GMC were perceived by focus groups (comprising doctors, nurses, allied health professionals, managers and patients) to meet adequate minimum language competence requirements. They recommended raising the existing cut scores, especially with respect to oral skills, while allowing some flexibility in assessing written skills. Pill and McNamara (2016) reported on a standard setting study to set a minimum standard of professionally relevant oral competence for three health professions, including medicine, as measured by the OET speaking subtest. McNamara et al. (2018) reported a collaboration between language testers, applied linguists and healthcare professionals to determine standards in interprofessional written communication about patient care.

Beyond domain relevance and cut-score setting, O'Neill, Buckendahl, Plake, and Taylor (2007) pointed out other important considerations that pertain when licensing bodies select and use ELP tests as proof of a doctor's proficiency. They observed that 'although in some contexts occupation-specific language proficiency tests for health professionals have been developed, factors such as frequency of test administrations and distribution of test centers worldwide

mean that less occupation-specific, international academic English language tests such as IELTS and TOEFL serve an important function in this regard' (ibid, p.297-298). Logistical factors such as test access, availability and security can be significant considerations for stakeholders such as medical licensing bodies who recruit internationally trained doctors from all over the world. However, such factors are often not readily acknowledged or appreciated by language testers and thus are rarely discussed in the literature. The policy and practice of peer or parallel professional organisations in other parts of the world can also be a factor which test stakeholders wish to consider.

The aim of this paper, therefore, is to describe ELP tests that were, or could be, considered by professional organisations as acceptable for medical registration purposes and to understand the differences between them, as well as the complex criteria a professional body may prioritise when selecting a test. Although the original study was completed in 2015, the paper takes account of subsequent changes in policy and practice. The discussion below is guided by three research questions:

RQ1. What criteria were considered essential by one major UK-based regulatory body when selecting an ELP test to evidence doctors' writing proficiency for professional registration and licensing?

RQ2. Which ELP tests have been selected by other regulatory bodies for the health professions worldwide to confirm an acceptable level of writing proficiency prior to registration?

RQ3. What validity evidence can be assembled to support the selection of an ELP writing test (or tests) for licensing purposes, and how can this be done?

3. Method

The study adopted a mixed methods approach involving a documentary audit and review of tests combined with desk-based internet research; follow-up contact with specific test providers was also undertaken. The study had three stages: Stage 1 - an initial audit of available tests of English language skills in order to determine the GMC's essential criteria for test selection for subsequent scrutiny; Stage 2 - an audit of ELP tests accepted by other medical and health-care professional bodies worldwide; and Stage 3 - expert scrutiny of selected tests. These three stages will be described below. Although the original study focused on all components of a given test (i.e. all four language skills), the remainder of this paper will relate specifically to the writing assessment components of the selected tests.

3.1 Stage 1 (an audit of available tests of English language skills to determine the GMC's essential criteria for preliminary test selection for later scrutiny)

This entailed conducting an audit of tests of English language writing skills potentially available for demonstrating English language proficiency prior to entry to a technical and high-risk profession. The aim was to select those which met essential quality criteria for possible consideration within UK GMC registration procedures.

To compile the list of available ELP tests, the first step was to establish criteria for selection in close consultation with the regulatory body concerned. At this stage, the process highlighted a range of key considerations for the GMC, in addition to matters of authenticity and domain specificity. The criteria agreed at the time of the study in 2014 included:

- (1) English language proficiency tests for which the GMC had already received requests for acceptance from overseas qualified doctors;

- (2) English language proficiency tests which were approved as a Secure English Language Test (SELTs) by the UK Home Office and its Visa and Immigration service (UKVI);
- (3) English language proficiency tests widely used in countries from which the GMC received a high volume³ of applications for registration; or
- (4) English language proficiency tests which were *medically* oriented.

As a result, the audit identified a total of 45 ELP tests. To filter down this preliminary list so as to select out those tests which represented genuine potential to be accepted as adequate evidence of a doctor's English proficiency level, the next step was to establish essential test quality selection criteria (in response to RQ1). An initial set of 13 criteria was proposed by the researchers. After review and discussion with the GMC team, eight essential quality criteria were prioritised. The results are reported in Section 4.1.

3.2 Stage 2 (identifying tests used for professional registration purposes worldwide)

This stage entailed desk research to investigate what regulatory and licensing bodies for medical and other healthcare professionals demanded as sufficient evidence of the level of English writing skills needed by applicants prior to formal registration and subsequent practice.

First, a list of regulatory bodies for healthcare professionals in the UK and internationally was compiled. The GMC requested that other healthcare professionals, such as dental, chiropractic, optical, osteopathic, pharmaceutical, were included alongside the medical profession. The desk-based audit was initially informed by previous studies of a similar nature (Merrifield, 2008, 2011), supplemented with a follow-up exercise to identify as many additional relevant organisations as possible. Organisations in major English-speaking countries including UK, Australia, New Zealand, USA, Canada, Ireland, South Africa, Hong Kong (China) and Singapore were first surveyed. At the request of the GMC, regulatory authorities in countries which were significant source countries for overseas-qualified doctors applying to the GMC for registration in the UK were added. This included Bangladesh, Egypt, India, Iraq, Nigeria, Sudan, Sri Lanka and Pakistan⁴. Further efforts were made to identify policy and practice regarding any language entry requirements imposed by regulatory authorities in EEA countries which were significant source countries for overseas-qualified doctors applying to the GMC: Bulgaria, Czech Republic, France, Germany, Greece, Hungary, Ireland, Italy, Lithuania, Malta, Netherlands, Poland, Portugal, Romania, Slovakia and Spain⁵. However, with the exception of Malta and Ireland, most of these EEA countries did not require evidence of English or any other language proficiency for medical registration prior to entry.

Following compilation of an extensive (though admittedly not exhaustive) list of almost 100 regulatory bodies for medical and other healthcare professionals in the UK and internationally, a systematic search of each regulatory body's website was undertaken to gather relevant information concerning English language proficiency requirements for their professional registration purposes for internationally qualified applicants, specifically any mention of accepting test scores from approved ELP tests; various search criteria were used to seek out

³ The threshold for country inclusion was agreed as 50 or more applications annually, according to 2008-2013 data routinely gathered by the GMC.

⁴ An appropriate medical regulatory authority could not be identified for all of these countries via the web-based exercise at the time of the study.

⁵ For most of these countries, it proved extremely difficult to identify relevant professional organisations for these countries mainly because the websites were in the language(s) of the country concerned; no translation capabilities or resources had been factored into this study.

this information, including key words such as: *(English) language (proficiency) requirements, registration (procedures/standards), overseas/international applicants, overseas qualification/assessment*, and similar relevant terms. Specific attention was also paid to record any mention of recent changes in policy and practice, or likely revision in the future. Once the relevant data had been gathered, the information was reviewed and analysed to identify any major trends and patterns. Particular attention was paid on how far the use of ELP requirements within registration procedures for healthcare professionals varied among UK-based and international regulatory bodies as well as across differing healthcare professions, and how far requirements for registration purposes aligned with existing GMC policy and practice. The results are presented in Section 4.2.

3.3 Stage 3 - Expert scrutiny of shortlisted writing tests

Taking the GMC's recognition of IELTS Band 7/7.5 as the accepted benchmark at the time of the research, Stage 3 was to determine whether the shortlisted writing tests (see Section 4.1) could reasonably be considered comparable to those provided by the IELTS writing component, and thus acceptable as evidence for demonstrating a doctor's English proficiency level. It is important to stress that the purpose of this study was not to examine (or critique) the authenticity of IELTS writing in terms of its match (or otherwise) to the TLU domain. IELTS was already accepted by the GMC and Berry et al (2013) had also confirmed that the test should be retained as an appropriate test of the English language competence of overseas-trained doctors. This is not to say, however, that further research to explore the match of IELTS to the medical TLU would not be enlightening. The specific purpose of this study was to examine the subtests of other ELP assessments which the GMC might need/wish to consider alongside IELTS as part of their registration and licensing procedures and to compare these systematically against the existing benchmark of IELTS. At the level of the writing subtest, this involved a comparative analysis of the contextual, cognitive and scoring features of 12 shortlisted writing tests.

To achieve this, we needed a detailed conceptual framework to help us account for the core component of the writing construct. A socio-cognitive framework (SCF) (Weir, 2005) was used to evaluate claims regarding the quality of the shortlisted tests on three core dimensions of test validation. The component of **context validity** concerns the contextual parameters of a task, which are often socially or externally determined in terms of the demands of the task setting, with its specified input and expected output. The component relating to **cognitive validity** concerns the individual language user and their cognitive processing to complete a task. **Scoring validity**, i.e. how the task performance is evaluated, is the component which combines with cognitive and context validity in an interactive, symbiotic relationship to constitute the overall construct validity of any test. The SCF had been successfully applied to many large-scale English language proficiency tests for purposes of test validation and future development (for example see Chan, Inoue and Taylor, 2015; Shaw and Weir, 2007; O'Sullivan, 2015).

In this study, we focused on relevant parameters from the SCF to evaluate each shortlisted writing test in relation to considerations which were regarded essential by a medical regulatory body when approving an acceptable test for registration purposes. These considerations included not only those relating to authenticity or domain specificity but also key practicality aspects such as test security, score integrity and dependability of results. A template of the test scrutiny form is provided in Appendix A.

To evaluate the task features of the writing tasks, we referred to the contextual parameters, i.e. the performance conditions or task features, under which a candidate produces the writing. Each shortlisted writing test was analysed in relation to 13 contextual parameters which have been established in the literature that are likely to influence writing performance (for example, see Chan, 2018; Shaw and Weir, 2007; Weigle, 2002; Weir, 2005). The parameters included: *number of tasks, response format and genre, task description, source texts (if any), domain, topic, communicative purpose, knowledge of criteria, writer-reader relationship, timing, text length, weighting and delivery mode*. We also examined the linguistic demands of the writing output expected from the test takers. Parameters included *discourse mode* which test takers are required to perform, and the requirement of *lexical level* (e.g. *proportion of frequency words, academic words*), *syntactic complexity, grammatical accuracy and coherence and cohesion* of these texts, see Table 1.

[insert Table 1 here]

To examine the assumed writing skills being assessed, we drew upon previous research studies and/or technical report on the cognitive validity of each shortlisted writing test. We completed the cognitive parameters in relation to Shaw and Weir's (2007) model of writing (see Table 2), which views writing activity as consisting of five distinct levels of processing. When integrated tasks were concerned, model of discourse synthesis (Spivey, 1990) which involves processing of connecting, selecting, and organising ideas from sources was also consulted.

[insert Table 2 here]

Lastly, we examined parameters in relation to the scoring dimensions of test validity: the *scoring approach, scoring criteria and reliability* of each writing. Further information on these parameters can be found in a worked example of analysis in Appendix 3.

The comparative analysis across selected writing tests was undertaken by a team of three writing assessment specialists to scrutinise the validity of the shortlisted tests as well as to evaluate essential practicality issues. All three experts had a PhD in Language Assessment and, at the time of the study, they had between five- and 15-years' experience of developing and validating writing tests.

The procedure began with a comprehensive literature search for published information about each writing test scrutinised, in particular for independent research studies where available, relating to all aspects of the test quality in question, in order to gather information for each aforementioned parameter (see Tables 1 and 2) in the test scrutiny template (see Appendix 1). Five samples (i.e. 5 parallel forms of the test tasks) for each shortlisted test were examined to verify the published claims, or to obtain more detailed information. The three experts drew on published descriptions of each writing test in order to complete the sections on Task Description, Timing, and so on, and checked these against the samples for each test. Previous research studies on the complexities of each test, where available, were cited. Objective data (e.g. number of tasks) and expert judgements (e.g. domain) on each of these parameters was entered to the template by the team, which was then completed in turn for each of the writing tests examined (as a worked example, see a completed scrutiny form for IELTS writing in Appendix 3).

After the test scrutiny form had been completed for each of the writing tests (see Appendix 3 for the worked example for IELTS), a focus group seminar was conducted. The team of three writing experts reviewed all the information for each test in each category to come to a consensus regarding the difficulty level of the test in relation to each parameter in comparison with the IELTS writing test. For example, the Certificate of Proficiency (CPE) writing test was determined to be more demanding than IELTS in terms of output text length because IELTS

writing requires candidates to produce a total of 400 words on two tasks (at least 150 words on Task 1 and at least 250 words on Task 2), whereas candidates are required to produce at least 520 words on two writing tasks on CPE (240-280 words on Task 1 and 280-320 on Task 2). Another example is that IELTS was determined to be less demanding than the Test of English as a Foreign Language (TOEFL) iBT in terms of communicative purpose. While Task 2 (essay task) on both tests were largely comparable in terms of their productive demands, TOEFL iBT Task 1 was determined to be more demanding than IELTS Writing Task 1. The former requires candidates to summarise the points in a listening passage and explain how they relate to specific points in a reading passage, whereas the latter requires candidates to collate different pieces of information from diagrams or tables in order to describe information.

Each parameter in the template carried equal weight. The team of three experts first considered the number of parameters which were judged more demanding (>), less demanding (<), or comparable (=) to IELTS in relation to context validity (e.g. task setting, complexity of input and output) and cognitive validity (e.g. target processes). The team then considered the parameters in relation to scoring validity (e.g. reliability and scoring approach) and other practical and security issues. Drawing on the comparative analysis of individual parameters between IELTS and the shortlisted tests, the team then made a holistic judgement concerning the relative difficulty of each writing test as a whole, in comparison with IELTS. The results are presented in Section 4.3.

4. Results

In this section, we report the results in relation to each of the three Research Questions.

4.1. Essential test quality selection criteria from the perspective of one medical regulatory body (RQ 1)

As mentioned in Method, to select tests which represented genuine potential to be accepted as sufficient evidence of a doctor's English proficiency level, an initial set of 13 possible selection criteria was proposed by the researchers. After review and discussion with the GMC team, the following eight essential quality criteria were prioritised. The reasons for exclusion of the other five criteria are provided in Appendix 2.

1. the existence of high-level security procedures for test production and delivery;
2. the existence of a test score verification system that allows score users to check and confirm test results;
3. assessment of all four communicative skills;
4. parallel forms of the test available for all test components;
5. appropriate evidence of the reliability of the test and the dependability of its scores;
6. an approved test centre (or centre network) routinely monitored and inspected for quality assurance;
7. demonstrable evidence, e.g. publicly available technical or research reports, that the test measures at the CEFR C1 level and/or above; and
8. representation of an appropriate test domain (i.e. academic or professional rather than just general orientation)

These eight criteria reflected the essential test quality for the purpose of professional registration and licensing from the perspective of one medical regulatory body. As mentioned in Method, the audit initially identified 45 tests of English language skills worldwide which were potentially available for demonstrating English language proficiency. 14 of these tests targeted an ability level below the critical target level of C1 and five tests did not assess all four

communicative skills, so 19 tests were quickly excluded from any further consideration or analysis. This left 26 tests, all of which included a writing component, for closer examination. Table 3 presents a summary of the results from this analysis in which the eight criteria have been recast as questions.

[insert Table 3 here]

Using the list of criterial questions shown in Table 3, the 26 tests were subjected to a further filtering process in order to select tests that met the GMC's essential quality assurance criteria for regulatory purposes for medical doctors.

Four out of the 26 tests which appeared to meet all eight essential quality criteria were initially shortlisted for test scrutiny. As suggested by the GMC, a follow-up enquiry letter was sent by email to providers of six tests requesting specific information concerning one or more of the eight essential criteria (i.e. the outstanding information that the researchers were unable to locate through the desk research). Test providers were invited to supply the outstanding information, or to indicate where it could be located in the public domain, within 21 days. Providers of these six tests supplied documentation to show that their tests met the outstanding criteria. This led to a total of nine tests which met the essential quality assurance criteria and thus could be prioritised for close scrutiny (Stage 2).

- Cambridge English: Advanced (CAE)
- Cambridge English: Proficiency (CPE)
- Internet-based Test of English as a Foreign Language (TOEFL iBT)
- Pearson Test of English - Academic (PTE-A)
- Occupational English Test (OET)
- Examination for the Certificate of Proficiency in English (ECPE)
- General English Proficiency Test – Advanced (GEPT Advanced)
- International ESOL and International Spoken ESOL tests (IESOL and ISESOL) – Level 2
- International ESOL and International Spoken ESOL tests (IESOL and ISESOL) – Level 3

The results of the test scrutiny are reported in Section 4.3.

4.2 Range of tests accepted by other healthcare regulatory bodies worldwide (RQ2)

The websites of almost 100 UK-based and international regulatory bodies for the healthcare professions were searched to investigate what such bodies required at the time of the original study (2014) as appropriate evidence of the level of English language proficiency needed by applicants for professional registration. Drawing upon the detailed information gathered through this search, a number of key trends and patterns were observed as described below. We also include information about new tests which have been accepted since the original study to reflect current practice. It should be noted that this section reports on the acceptance of a given test (or tests) as a whole, rather than on the writing assessment component(s) in particular. Nevertheless, the information is included here to give a comprehensive picture of the wide variety of policy and practice which is implemented by healthcare regulatory bodies around the world when deciding on an acceptable test for registration purposes. This, in turn, may suggest some implications for decisions on approaches to assessing the writing proficiency of doctors and similar medical professionals, e.g. concerning the international transferability of ELP qualifications across medical jurisdictions.

In 2014, IELTS was the preferred ELP test for professional registration purposes across the healthcare professions in the UK. Among other tests mentioned were professional competency tests internal to the profession which might contain some sort of English language or communication skills element (e.g. Clinical OSCE). The General Osteopathic Council (GOsC) mentioned preferring IELTS but also accepted any test demonstrating that the applicant has met level C1 of the Common European Framework for Languages (CEFR). Some examples of these tests provided on their website included Cambridge English: Advanced (CAE) and TOEFL (The requirement was last updated in 2014; <https://www.osteopathy.org.uk/news-and-resources/document-library/registration/english-language-test-requirements/>).

For dental practitioners seeking admission to dental performer lists, which are administered by regional healthcare commissioning bodies throughout the UK, the General Dental Council required applicants from EEA countries and those who qualified outside the EEA to provide appropriate evidence that they have the necessary knowledge of English before they can be registered. In terms of the ELP test for registration, they accepted an overall of IELTS 7.0, with no individual bands lower than 6.5 (https://www.gdc-uk.org/docs/default-source/english-language-controls/gdc-guidance-on-english-language-controls.pdf?sfvrsn=88adf167_2).

In **Australia**, at the time of the study, the most commonly mentioned tests were IELTS and the Occupational English Test (OET), originally developed in and for Australia by the Language Testing Research Centre at the University of Melbourne in the late 1980s. Almost all the boards surveyed reported accepting both IELTS and OET scores. From 2015, the Australian Health Practitioner Regulation Agency (AHPRA) introduced revised standards to accept the PTE Academic and TOEFL iBT for ten ⁶ health-care professions (<https://www.ahpra.gov.au/Registration/Registration-Standards/English-language-skills.aspx>). The Australian Medical Council Inc additionally accepted the NZREX (New Zealand Registration Examination) and the GMC's PLAB Test, as does the Medical Board of Australia. From March 2019, the Nursing and Midwifery Board of Australia also accepted IELTS, OET, PTE Academic and TOEFL iBT. In December 2019, The Aboriginal and Torres Strait Islander Health Practice Board introduced a revised standard. It requires all applicants for registration must demonstrate their English language competency through the successful completion of an approved program of study, or holding a qualification that is at the Australian Qualifications Framework Level 4 or Certificate IV or higher. In **New Zealand**, all the boards surveyed accepted both IELTS and OET (standard approved in October 2010 and updated on March 2018) (<https://www.mcnz.org.nz/assets/Policies/ad921b8a7f/English-language-policy.pdf>).

It proved extremely difficult to track down a list of appropriate regulatory authorities in the **USA**, at either federal or state level, in order to identify what English language proficiency requirements were expected from International Medical Graduates (IMGs). Further work may be needed in this area using a different methodological approach to that adopted for investigating other parts of the English-speaking world. What seemed clear was that the United States Medical Licensing Examination (USMLE) incorporated within it an English language and communication skills component and it may be that this serves universally across the USA for registration purposes (inferred from the website of the Educational Commission for Foreign Medical Graduates). The use of IELTS was mentioned in relation to the International Commission on Healthcare Professionals.

When exploring policy and practice in **Canada**, it proved much easier to drill down from national to provincial level than had been the case for the USA. Several of Canada's provincial Colleges of Physicians and Surgeons accepted both IELTS and TOEFL iBT. Nursing-related

⁶ Chinese medicine practitioners, chiropractors, medical radiation practitioners, occupational therapists, optometrists, osteopaths, pharmacists, physiotherapists, podiatrists, and psychologists.

bodies mentioned using IELTS and also the Canadian English Language Assessment Benchmark Assessment for Nurses (CELBAN), an English language test developed specifically for the nursing profession in Canada. Regulatory bodies for pharmacists appeared to accept a wider range of tests in addition to IELTS and TOEFL iBT: the Canadian Test of English (CANTEST) and the Michigan English Language Assessment Battery (MELAB), though the latter has since been replaced by the Michigan English Test (MET).

No information could be found on ELP requirements for the **Caribbean**, though the Caribbean Association of Medical Councils (CAMC) examination appeared to be in English and claimed to test 'written and factual knowledge' as well as clinical communication and problem-solving'.

In **Ireland** the relevant boards accepted IELTS and OET, though the Irish Medical Council also accepted CAE and United States Medical Licensing Examination (USMLE) if all of part 1 and part 2 have been completed (<https://www.medicalcouncil.ie/registration-applications/first-time-applicants/english-language.html>). The Health Professionals Council of **South Africa** accepted IELTS but mentioned no other tests. In **Hong Kong** the Medical Council had its own Proficiency Test in Medical English which formed Part II of a 3-part Licensing Examination (https://leip.mchk.org.hk/EN/g_info.html). The test was a written English paper designed to ensure that the candidates' command of English is adequate for professional purposes.⁷ In **Singapore** IELTS, TOEFL iBT and OET were all accepted as providing evidence of English language proficiency (<https://www.healthprofessionals.gov.sg/smc/becoming-a-registered-doctor/documentation-requirements-processing-time>).

Unfortunately, no relevant information could be found when searching the websites of various healthcare regulatory bodies in countries which were significant source countries for IMGs seeking UK registration (Bangladesh, Egypt, India, Iraq, Nigeria, Sudan and Sri Lanka). The only exception was the **Pakistan** Medical and Dental Council which accepted both IELTS and TOEFL scores (though the latter score seemed to relate to the former TOEFL test).

No useful or accessible websites could be found when searching for information for EEA countries which were listed by the GMC as significant source countries for IMGs seeking UK registration. For most of these countries, it proved extremely difficult to reliably identify relevant professional organisations so as to determine policy and practice regarding certification of language proficiency at the national level.

Having reported the ELP tests accepted for evidence of English proficiency by professional registration bodies in the UK and elsewhere, we now report the results of the test scrutiny.

4.3 Test scrutiny – potential suitability for medical registration purposes (RQ3)

As reported in Section 4.1, nine tests were shortlisted for test scrutiny and analysed using the template in Appendix 1. The nine tests were analysed in two phases. In the first phase, five tests, i.e. the four tests (CAE, CPE, TOEFL iBT and PTE-A) which originally met the eight criteria, plus OET, were analysed. OET was prioritised due to its specifically medical orientation. The remaining four tests (ECPE, GEPT, IESOL/ ISESOL level 2 and IESOL/ ISESOL level 3), i.e. those that met the eight criteria after follow-up enquiries, were analysed in a follow-up phase. Exactly the same methodology was applied in the two phases, but limitations on space mean that only the test scrutiny results for CAE, CPE, TOEFL iBT, PTE-

⁷ Part I is an Examination in Professional Knowledge consisting of two papers, each comprising 120 multiple choice questions testing candidates' medical knowledge. Part III is a Clinical Examination designed to test ability to apply professional knowledge to clinical problems; it is possible that speaking and listening skills are evaluated in Part III of the Licensing Examination, although it should be noted that candidates can choose to respond in either English, Cantonese or Putonghua.

A and OET are reported here (results for the other four tests can be found in Taylor and Chan, 2015).

As described in Section 3.3, after the three researchers had completed the test scrutiny form (see Appendix 1) for each test, they reviewed the information for each test in each category and came to an overall expert judgement concerning the relative difficulty of the test in each of the categories when compared against the IELTS writing test. The focus group then drew on the individual analyses to form an overall evaluation of the relative difficulty of each writing test as a whole, in comparison with IELTS. A summary of the results is shown in Table 4.

[insert Table 4 here]

On balance, three of the writing tests (i.e. CPE, PET-A and TOEFL iBT) were judged to be more demanding than the IELTS writing test (which comprises an informational summary based on data provided and a discursive essay); CAE writing was judged as comparable to IELTS writing and OET writing as less demanding than IELTS writing. The analysis and reasoning underpinning this set of expert judgements is expanded below.

CPE is designed to target a *single proficiency level* - the C2 level of the CEFR; therefore, the writing test reflects tasks and skills appropriate to this single level (i.e. essay, article, letter, report or review). The CPE Writing test was judged by the expert focus group as more demanding overall than an IELTS Writing test in a number of respects. The CPE test has a more complex skills focus than IELTS, sampling writing skills more broadly: reading-into-writing and writing. Cognitive processing in CPE was judged to be more demanding due to the need for knowledge transformation from reading sources to writing; multiple verbal source texts in the task; and higher demands in organizing the text to fulfil the task purpose. The CPE was also more demanding in terms of topic (more specific content knowledge is required), purpose, timing and expected text length than the IELTS Writing. Most of the linguistic demands for CPE writing were considered by the expert judges more demanding than the IELTS Writing test.

Similar to IELTS, PTE-A is designed to target a *range of proficiency levels*; therefore, the writing test reflects tasks and skills appropriate across these multiple levels in a similar way to IELTS. The PTE-A Writing test was judged by the expert focus group to be slightly more demanding in terms of number of tasks, response format and genre, demand of source texts, timing and the text length of expected response. The PTE-A test also involves a wider range of response format and genre, such as *summarise written text*, *summarise spoken text* and *essay*. The integrated tasks require test takers to process written and spoken source materials under timed conditions, therefore adding an extra level of demand.

Similar to IELTS, TOEFL iBT is designed to target a range of proficiency levels; therefore, the writing test aims to reflect tasks and skills appropriate across these multiple levels in a similar way to IELTS. The TOEFL iBT Writing test (involving an integrated summary response to spoken and written input, plus an independent topic-based essay) is similar to IELTS in its demands, but was judged by the focus group to be more demanding due to a higher demand in source texts, timing and the text length of expected response. Both independent and integrated writing skills are tested. Test takers have to process written and spoken source materials in the integrated task.

CAE is designed to target a *single proficiency level* - the C1 level of the CEFR; therefore, the writing test reflects tasks and skills appropriate to this single level. The CAE Writing test was judged by the expert focus group to be more demanding than the IELTS test in terms of source texts and purpose. The CAE test also involves a broader skills focus, including reading-into-

writing and writing of a letter, a proposal, a competition entry, a contribution to a guidebook or a work-related task. However, candidates are able to choose 2 out of 4 writing options offered in Task 2, thus allowing them to select those topics/genres they find more accessible or which they perceive as easier. In addition, the CAE test is less demanding in terms of timing and text length. Considering all these parameters, the CAE Writing test was considered to be comparable to the IELTS Writing test. All linguistic demands imposed by the CAE writing test were considered to be comparable to those of the IELTS Writing Test.

OET⁸ is designed to target *a range of proficiency levels*; therefore, the writing test, at the time of the original study, reflected tasks and skills appropriate across these multiple levels in a similar way to IELTS. On the one hand, the OET Writing test was judged by the expert focus group to be more demanding than IELTS in terms of candidates having to read more source texts, write on a more specific topic and address a tightly defined writer-reader relationship. However, the OET Writing test has only one task (a letter of referral, transfer or discharge) and the skills focus is targeted specifically on writing a letter based on medically contextualized information which is provided to candidates who already have a medical background. Most linguistic demands, except discourse mode, imposed by the OET writing test were considered to be comparable to the IELTS Writing Test (Band 7 or above). Considering all these aspects, the OET Writing was considered to be a less demanding test overall than IELTS Writing given that it involves only one task and is taken by candidates who already possess a healthcare background which potentially gives them an advantage.

5. Discussion

5.1 Comparing policy and practice in overseas doctors' registration and licensing

The findings suggested that, despite some variations, **policy and practice for overseas doctors' registration** in the UK was comparable to many other international medical licensing bodies. At the time of writing, the GMC continues to set the ELP requirements for overseas-qualified applicants at an overall IELTS Band 7.5. A minimum of Band 7.0 is required in the writing subtest (as for each of the skill subtests in IELTS). Other UK medical bodies⁹ set a comparable benchmark to the GMC, or adopted the GMC standard (i.e. 7.5 overall with a minimum of 7.0 in writing). Based on the survey of regulatory boards' policy and practice (see Section 4.2), this requirement appears to sit at the higher, more demanding end of the scale. Healthcare bodies in Australia, Ireland, Canada and Singapore appeared to accept a slightly lower IELTS standard on balance, i.e. an overall 7.0 with a minimum 6.5 or 7.0 in each subtest component. In New Zealand, 7.0 was required for the IELTS Reading and Writing subtests but 7.5 for Listening and Speaking. Comparable OET score levels accepted by organisations in Australia, New Zealand and Ireland were a minimum of B for each component (N.B. OET does not report an overall grade for the test as a whole). The equivalent score required on TOEFL iBT appeared to be generally 94-96 overall with a minimum of 24-27 in writing. In South Africa an overall band 6.0 on IELTS appeared to be the standard, but with no mention of minimum component scores.

Regarding the **range of ELT tests accepted**, it appears that IELTS remains the preferred ELP test for professional registration purposes across the healthcare professions in the UK. The General Osteopathic Council (GOsC) in the UK mentioned preferring IELTS but also

⁸ OET was undergoing a revision during the time of the research. For the latest format of OET, please refer to <https://www.occupationalenglishtest.org/test-information/writing/>

⁹ The bodies include Royal College of Anaesthetists, Royal College of Obstetricians and Gynaecologists, Royal College of Ophthalmologists, Royal College of Paediatrics and Child Health, Royal College of Ophthalmologists, Royal College of Pathologists and Royal College of Psychiatrists

accepted any test demonstrating that the applicant has met meet level C1 of the Common European Framework for Languages (CEFR). Following the study, in 2018, the GMC opted to accept the Occupational English Test (OET) as proof of a doctor's language competency, as an alternative to IELTS. In Australia, New Zealand and Ireland, both IELTS and the OET were widely recognised and accepted. But as described in Section 4.2, since the study, several healthcare regulatory bodies have accepted a wider range of ELP tests. For example, PTE-A and TOEFL iBT were accepted by the AHPRA in Australia the from 2015. In general, there seems to be a tendency towards widening the range of ELP tests accepted by these international medical licensing bodies. The need to address the issue of test comparability, therefore, becomes even more pressing.

5.2 Determining the comparability of tests

As labour mobility in certain professions like the healthcare professions is increasing worldwide, language test comparability has become an issue especially in the context of visa application, educational admission or professional registration. Questions of test equivalence are perhaps best explored through an empirical experimental study in which a carefully selected sample of representative test-takers is required to sit all the different tests under consideration in order that their scores on each of the tests can be compared against one another as well as the predictive power of each test score in relation to test takers' written performance in the real-life healthcare context. In practice, of course, this approach is rarely feasible, and though it may be possible when comparing just two tests, it is very unlikely to be a realistic proposition when seeking to compare three or more tests. Even if it were logistically possible, such an approach risks neglecting important cognitive and contextual validity features of the tests under scrutiny. Among other things, it would be difficult to take account of factors such as candidates' test familiarity and level of test preparation. For this reason, an alternative, more content-analytic approach was necessary as well as desirable in this study, drawing on (i) information and evidence relating to the perceived cognitive and contextual validity of the test content, and (ii) information concerning score interpretation, value and comparability provided by test providers themselves or publicly available in independent research literature.

Weir (2005) argued that social moderation and calibration might only address part of the test comparability issue. Language tests are 'simply not comparable' (ibid, p.296) unless they exhibit equivalence in terms of key validity parameters and unless this can be demonstrated through empirical evidence and cogent argumentation by those who are well qualified to adjudicate. Through the methodology of this study, we hope to demonstrate the value of working closely with the regulatory body, using a tried and tested analytical framework, to take account of the criteria they consider are essential for their registration and licensing purposes. Such criteria might not be the same (and may even be at odds with) those criteria typically prioritised by language testers. Nevertheless, it is surely important to acknowledge the perspectives of all relevant stakeholders in the test and to incorporate such considerations into practical frameworks for test selection and validation.

6. Conclusion

6.1 Limitations and directions for future research

This study sought to address a set of complex questions concerning the comparability of ELP tests to inform policy and practice for overseas doctors' registration. A number of caveats should be borne in mind when reviewing the survey of policy and practice that was undertaken as part of this study and when seeking to interpret the findings outlined above.

While the research team sought all relevant information as thoroughly and systematically as possible within the scope of the investigation, it is important to acknowledge that additional relevant information may be in existence; the fact that this was not readily available or could

not be found by the research team may affect the analyses. It is not easy to be definitive about the current state of affairs for any given regulatory body since policy and practice change over time, e.g. as approved tests are revised, as new tests become available, as test recognition and currency arrangements are reviewed. Though some bodies claim regular review and updating of their policy and practice in this area, their websites do not always appear to bear this out. We have updated some key changes in policy where we are aware of these, e.g. the recognition of TOEFL iBT and PTE-A by AHPRA in Australia and the recognition of OET by the GMC. However, it may be that policy and practice for other professional bodies or geographical regions has also changed in recent years.

There may be important differences to note between the nature and status of different professional healthcare organisations: some have statutory and/or regulatory powers, while others seem to be more like professional membership bodies or collegial societies with advisory rather than regulatory power. Regulatory bodies' policy on eligibility for exemption from having to take an ELP test appears to vary substantially; exemption may be on grounds of successful secondary education in English, or tertiary qualifications taught and assessed in English in listed English-speaking countries, or some other combination of circumstances. Some boards make mention of accepting other ELP tests from time to time but give no further details on this. In some parts of the world, e.g. the USA, it can be difficult to discern the regulatory relationship between the national/federal level and the state/provincial level.

As explained in Section 3.3, this paper does not directly address the relevance of a writing test and tasks to the target language use (TLU) domain, i.e. the professional medical context, in terms of task authenticity or cognitive validity. Nor is it concerned with the determination or setting of standards of minimum competence in written communication skills. The important contribution of domain experts in both of these areas is well recognised, as evidenced by Berry et al (2013) and as discussed by Manias and McNamara (2015) and McNamara et al (2018). We neither ignore nor underestimate the importance of questions concerning the issue of 'domain relevance' or 'domain specificity'. Further research is undoubtedly needed to examine more closely the level of match between any of the writing subtests mentioned above and the professional TLU domains with which they are associated to confirm their predictive power within the context of concern. Linked to this, it can be argued, is the complex issue of test washback and impact. Though hard to research and evaluate, washback has certainly been a consideration in test design for several decades now. It is reasonable to assume that a test which encourages people to prepare themselves for the domain in which they will work may end up having greater predictive power than one which is less well targeted to the intended context in terms of its task demands, though this remains an empirical question. Set against this, of course, is the practical and theoretical challenge for test designers, and indeed for domain experts themselves, of how to specify the domain in terms of its distinctive or essential contextual and cognitive parameters. This process is likely to be as challenging for those developing or selecting ELPs to meet the needs of the international medical and healthcare world as it is for those responding to the growing diversity of academic subject fields in higher education. An added challenge is the 'fluid' nature of much academic and professional life today which can mean that the expected TLU domain may not prove to be as stable or constrained as test designers might hope? While acknowledging all these issues, our primary focus and concern in this paper has been to try and shed light on some of the other key considerations facing medical and other healthcare regulatory bodies as they have to make real-world decisions about selecting and using ELPs as part of their professional licensing procedures.

Despite their limitations, studies such as the one described in this article can hopefully play an important role in assisting health-care regulatory and licensing organisations to make informed decisions regarding which language test(s) to use as acceptable evidence of overseas-trained doctors' English proficiency. It is our hope that this article will be of interest and value to other researchers, language testers, and policy-makers worldwide – both in terms of the research questions it raises and the methodological approach it offers.

References

- Chan, S. (2018). *Defining Integrated Reading-into-Writing Constructs: Evidence at the B2 C1 Interface. The English Profiles Series 08*. Cambridge University Press.
- Chan, S., Inoue, C., & Taylor, L. (2015). Developing rubrics to assess the reading-into-writing skills: A case study. *Assessing Writing*, 26, 20–37.
- Berry, V., O'Sullivan, B., & Rugea, S. (2013). Identifying the appropriate IELTS score levels for IMG applicants to the GMC register. *GMC Online Research Reports*, 1–150.
- Elder, C., McNamara, T., Woodward-Kron, R., Manias, E., McColl, G., Webb, G., & Pill, J. (2013). Towards improved healthcare communication: Development and validation of language proficiency standards for non-native Englishspeaking health professionals (Final report for the Occupational English Test Centre). Melbourne: The University of Melbourne, Language Testing Research Centre.
- Khalifa, H., & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading, Studies in Language Testing 29*. Cambridge: Cambridge University Press.
- Knoch, U., McNamara, T., Woodward-Kron, R., Elder, C., Manias, E., Flynn, E., & Zhang, Y. (2015). Towards improved language assessment of written health professional communication: the case of the Occupational English Test. *Papers in Language Testing and Assessment*, 5(2), 60–66.
- Knoch, U. Elder, C. Flynn E, Manias, E., McNamara, T. & Woodward-Kron, R. (2018) Operationalizing an expanded construct of writing proficiency for an occupation-specific language test: from workplace criteria to a workable rating scale. Presented as part of invited Colloquium: Crossing Disciplinary Boundaries: Challenges to Construct Definition in ESP/LSP Assessment. Applied Linguistics Association of Australia Conference, Wollongong, November.
- Manias, E., & McNamara, T. (2015). Standard setting in specific-purpose language testing: What can a qualitative study add? *Language Testing*, 33(2), 235–249.
- McNamara, T., Elder, C., Flynn, E., Knoch, U., Manias, E., Woodward-Kron, R., Yahalom, S. (2018). Collaboration in the design of specific purpose language tests in the health setting. *Applied Linguistics and Professional Practice* 13 (3), p.189-210.
- Merrifield, G. (2007). *An impact study into the use of IELTS by professional associations and registration entities: Canada, the UK and Ireland. IELTS Research Reports* (Vol. 11).
- Merrifield, G. (2008). An impact study into the use of IELTS as an entry criterion for professional associations – Australia, New Zealand and the USA. *IELTS Research Reports*, 8, 1–43.
- O'Neill, T. R., Buckendahl, C. W., Plake, B. S., & Taylor, L. (2007). Recommending a Nursing-Specific Passing Standard for the IELTS Examination. *Language Assessment Quarterly*, 4(4), 295–317. <https://doi.org/10.1080/15434300701533562>
- O'Sullivan, B. (2015). Aptis Test Development Approach. *Aptis Technical Reports*, 2015(1), 1–19.
- Pill, J. (2015). Drawing on indigenous criteria for more authentic assessment in a specific-purpose language test: Health professionals interacting with patients. *Language Testing*, 33(2), 175–193. <https://doi.org/10.1177/0265532215607400>
- Pill, J., & McNamara, T. (2015). How much is enough? Involving occupational experts in setting standards on a specific-purpose language test for health professionals. *Language Testing*, 33(2), 217–234. <https://doi.org/10.1177/0265532215607402>
- RAND. (2009). International Comparison of Ten Medical Regulatory Systems. Technical report for the GMC. Retrieved from http://www.gmc-uk.org/International_Comparison_of_Ten_Medical_Regulatory_Systems___final_report.pdf_25404378.pdf
- Shaw, S., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing, Studies in Language Testing 26*. Cambridge: Cambridge University Press and Cambridge ESOL.
- Spivey, N. (1994). Discourse synthesis: Constructing texts in reading and writing. Outstanding Dissertation Monograph. Newark, DE: International Reading Association.

- Taylor, L., & Chan, S. (2015). IELTS Equivalence Research for the purpose of GMC Registration. *GMC Research Report*, 1–279.
- Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*, 22(3), 1–20.
- Wette, R. (2011). English proficiency tests and communication skills training for overseas-qualified health professionals in Australia and New Zealand. *Language Assessment Quarterly*, 8(2), 200–210.

Appendix 1. Test scrutiny template (blank)

TEST:				COMPONENT:	Writing
Task Setting					
Number of tasks					
Response format and Genre					
Task description					
Source texts					
Domain					
Topic (Content knowledge)					
Communicative purpose					
Knowledge of criteria					
Writer-reader relationship					
Timing					
Text length of expected response					
Weighting					
Delivery mode					
Cognitive dimension					
Skills assessed					
Level of cognitive processing					
Linguistic demands					
Discourse mode (rhetorical task)	Descriptive	Narrative	Expository	Argumentative/persuasive	Instructive/Referral
Lexical complexity	Word frequency				
	Ratio of types to tokens				
	Lexical density				
Syntactic complexity					
Grammatical accuracy					
Coherence and cohesion					
Scoring					
Scoring approach					
Scoring criteria					
Reliability					
SEM					

Appendix 2. Criteria excluded in the ELP test selection and reasons for exclusion

Five of the original 13 criteria were de-prioritised in discussion with the GMC as explained below (not in order of importance):

- *inclusion of a direct face-to-face speaking test component* (rather than just a semi-direct test of oral proficiency which assesses test takers' oral proficiency in a non face-to-face setting without a human interlocutor, e.g. via a language laboratory or computer)
 - while a face-to-face speaking test (such as exists in IELTS) may be preferable in some respects because it permits the testing of reciprocal interaction, there exist some well-researched and well-regarded ELP tests that adopt a semi-direct format (e.g. TOEFL iBT); since some of these are already accepted for international professional registration purposes, it seemed inappropriate to exclude these tests at an early stage
- *formal acceptance/approval/recognition of the test by comparable registration bodies for medicine and similar professions*
 - although the GMC was keen to be aware of what other professional registration bodies do, it did not wish to be restricted by their policy and practice
- *geographical and temporal accessibility of the test* (i.e. internationally available on a regular basis)
 - although this criterion had been of importance in previous studies on acceptable English language proficiency tests undertaken medical registration bodies (e.g. O'Neill et al, 2007), accessibility was considered by the GMC to be less relevant to its immediate needs given that the primary concern was not 'test reach' for recruitment purposes
- *easy availability of comprehensive and up-to-date information for relevant stakeholder groups, i.e. test takers, test score users*
 - although the provision of essential published information concerning test purpose and quality is generally regarded as an important measure of test quality, this criterion was considered to be redundant if the other criteria had been met
- *evidence of a commitment to comprehensive on-going research for purposes of test validation and improvement*
 - while it is generally accepted among language testers that a test should have a sound basis in relevant theory and empirical research, the GMC believed that recently developed tests which had not yet had time to develop an extensive research and validation literature for themselves should not necessarily be excluded from consideration

Appendix 3. Worked example of test scrutiny template showing results for IELTS Writing

TEST:	International English Language Testing System (IELTS)	COMPONENT:	Writing
Task Setting			
Number of tasks	Two (Task 1 and Task 2)		
Response format and Genre	Task 1: Continuous writing; Information transfer from multiple non-verbal inputs Task 2: A longer piece of continuous writing; Essay		
Task description	<p>The IELTS Writing test is designed to assess a wide range of writing skills, including how well candidates:</p> <ul style="list-style-type: none"> • write a response appropriately • organise ideas • use a range of vocabulary and grammar accurately <p>In Task 1 candidates are presented with a graph, table, chart or diagram. Candidates are asked to describe, summarise or explain the information in their own words. This might involve describing and explaining data, describing the stages of a process or how something works, or describing an object or event.</p> <p>In Task 2 candidates are asked to write an essay in response to a point of view, argument or problem. Students should find the issues interesting and easy to understand.</p>		
Source texts	N/A		
Domain	Social Academic		
Topic	The issues raised are of general interest to, suitable for and easily understood by candidates entering undergraduate or postgraduate studies or seeking professional registration.		
Purpose	<p>Task 1:</p> <ul style="list-style-type: none"> • To transferring informing from multiple inputs • To collate different pieces of information in order to describe, summarise or explain the information <p>Task 2:</p> <ul style="list-style-type: none"> • To write a persuasive essay • To defend or attack a particular argument or opinion, compare or contrast aspects of an argument, and give reasons for the argument. 		
Knowledge of criteria	<p>Publish information about how the tasks are scored, including band descriptors for individual tasks, and procedures for scoring, are provided on the official website (http://takeielts.britishcouncil.org/prepare-test/how-prepare-ielts-test) as well as in a booklet <i>Information for candidates</i> (http://takeielts.britishcouncil.org/sites/default/files/Information_for_Candidates_booklet.pdf)</p>		
Writer-reader relationship	Candidates are expected to write within a formal context and to demonstrate sensitivity to their audience. However, a particular audience is not always specified in either Task 1 or Task 2.		
Timing	The IELTS Writing test takes 60 minutes. Students should spend 20 minutes on Task 1, and 40 minutes on Task 2. Students need to manage their own time.		
Text length	Task 1: at least 150 words Task 2: at least 250 words		
Weighting	Task 2 is worth twice as much as Task 1 in the IELTS Writing test.		
Mode	Mainly paper-based		
Skills assessed	In both tasks, candidates are assessed on their ability to write a response which is appropriate in terms of content, the organisation of ideas, and the accuracy and range of vocabulary and grammar.		

Cognitive processing	<p>Task 1:</p> <ul style="list-style-type: none">• Macro-planning: goal setting and task representation• Transforming ideas from non-verbal inputs• Organising ideas• Translating• Micro-planning• Monitoring and revision <p>(Bridges, 2010)</p> <ul style="list-style-type: none">• Comprehending non-graphic task instructions• Comprehending (and interpreting) the components of graphs• Re-presenting or re-producing the non-graphic and graphic information as continuous discourse in written form in English as a foreign language <p>(Yu, Rea-Dickins and Kiely, 2011)</p> <p>Task 2:</p> <ul style="list-style-type: none">• Macro-planning: goal setting and task representation• Generating ideas• Organising ideas• Translating• Micro-planning• Monitoring and revision <p>(Weir, O'Sullivan, Jin and Bax, 2007)</p>																																																																																				
Linguistic demands																																																																																					
Discourse mode	Descriptive Expository Argumentative/persuasive																																																																																				
Lexical Level	<p>From test provider</p> <p>To achieve Band 7, candidates are expected to use a sufficient range of vocabulary to allow some flexibility and precision and use less common lexical items with some awareness of style and collocation.</p> <p>From research studies</p> <p>Word frequency (calculates the percentage of words in the scripts at each band level that fell into the first 1000 (K1) and 2000 (K2) most frequently used English words, and those that were found in the Academic Word List (AWL))</p> <p>Task 1</p> <table><tr><th rowspan="2"></th><th colspan="3">L1 Chinese</th><th colspan="3">L1 Spanish</th></tr><tr><th>K1</th><th>K2</th><th>AWL</th><th>K1</th><th>K2</th><th>AWL</th></tr><tr><td>Band 7</td><td>65.38</td><td>9.58</td><td>13.81</td><td>60.21</td><td>10.26</td><td>17.62</td></tr><tr><td>Band 8</td><td>-</td><td>-</td><td>-</td><td>65.54</td><td>7.57</td><td>15.60</td></tr></table> <p>Task 2</p> <table><tr><th rowspan="2"></th><th colspan="3">L1 Chinese</th><th colspan="3">L1 Spanish</th></tr><tr><th>K1</th><th>K2</th><th>AWL</th><th>K1</th><th>K2</th><th>AWL</th></tr><tr><td>Band 7</td><td>57.69</td><td>13.59</td><td>13.91</td><td>56.94</td><td>10.85</td><td>18.51</td></tr><tr><td>Band 8</td><td>-</td><td>-</td><td>-</td><td>70.61</td><td>7.27</td><td>13.48</td></tr></table> <p>(Banerjee, Franceschina & Smith, 2007)</p> <p>Tokens (the total number of words) and Types (the total number of different word forms)</p> <table><tr><th rowspan="2"></th><th colspan="2">Task 1 Means (SD)</th><th colspan="2">Task 2 Means (SD)</th></tr><tr><th>Tokens</th><th>Types</th><th>Tokens</th><th>Types</th></tr><tr><td>Band 7</td><td>223.6 (51.8)</td><td>94.2 (20.6)</td><td>312.2 (56.8)</td><td>159.0 (25.6)</td></tr><tr><td>Band 8</td><td>230.6 (39.1)</td><td>102.6 (26.5)</td><td>323.3 (31.5)</td><td>160.4 (12.9)</td></tr></table> <p>(Banerjee, Franceschina & Smith, 2007)</p> <p>Ratio of types to tokens (This is a measure of lexical variation/diversity because the higher the number of types in relation to tokens the more varied/diverse the vocabulary used.)</p> <table><tr><th rowspan="2"></th><th colspan="2">TTR</th></tr><tr><th>Task 1</th><th>Task 2</th></tr><tr><td>Band 7</td><td>42.9 (6.6)</td><td>51.3 (5.1)</td></tr><tr><td>Band 8</td><td>44.7 (9.0)</td><td>49.8 (3.5)</td></tr></table> <p>(Banerjee, Franceschina & Smith, 2007)</p>		L1 Chinese			L1 Spanish			K1	K2	AWL	K1	K2	AWL	Band 7	65.38	9.58	13.81	60.21	10.26	17.62	Band 8	-	-	-	65.54	7.57	15.60		L1 Chinese			L1 Spanish			K1	K2	AWL	K1	K2	AWL	Band 7	57.69	13.59	13.91	56.94	10.85	18.51	Band 8	-	-	-	70.61	7.27	13.48		Task 1 Means (SD)		Task 2 Means (SD)		Tokens	Types	Tokens	Types	Band 7	223.6 (51.8)	94.2 (20.6)	312.2 (56.8)	159.0 (25.6)	Band 8	230.6 (39.1)	102.6 (26.5)	323.3 (31.5)	160.4 (12.9)		TTR		Task 1	Task 2	Band 7	42.9 (6.6)	51.3 (5.1)	Band 8	44.7 (9.0)	49.8 (3.5)
	L1 Chinese			L1 Spanish																																																																																	
	K1	K2	AWL	K1	K2	AWL																																																																															
Band 7	65.38	9.58	13.81	60.21	10.26	17.62																																																																															
Band 8	-	-	-	65.54	7.57	15.60																																																																															
	L1 Chinese			L1 Spanish																																																																																	
	K1	K2	AWL	K1	K2	AWL																																																																															
Band 7	57.69	13.59	13.91	56.94	10.85	18.51																																																																															
Band 8	-	-	-	70.61	7.27	13.48																																																																															
	Task 1 Means (SD)		Task 2 Means (SD)																																																																																		
	Tokens	Types	Tokens	Types																																																																																	
Band 7	223.6 (51.8)	94.2 (20.6)	312.2 (56.8)	159.0 (25.6)																																																																																	
Band 8	230.6 (39.1)	102.6 (26.5)	323.3 (31.5)	160.4 (12.9)																																																																																	
	TTR																																																																																				
	Task 1	Task 2																																																																																			
Band 7	42.9 (6.6)	51.3 (5.1)																																																																																			
Band 8	44.7 (9.0)	49.8 (3.5)																																																																																			

	<p>Lexical density (This measure compares the number of lexical words used to the number of grammatical words.)</p> <table><tr><td></td><td colspan="2">Lexical density</td></tr><tr><td></td><td>Task 1</td><td>Task 2</td></tr><tr><td>Band 7</td><td>0.82 (0.15)</td><td>0.75 (0.15)</td></tr><tr><td>Band 8</td><td>0.82 (0.21)</td><td>0.69 (0.13)</td></tr></table> <p>(Banerjee, Franceschina & Smith, 2007)</p>		Lexical density			Task 1	Task 2	Band 7	0.82 (0.15)	0.75 (0.15)	Band 8	0.82 (0.21)	0.69 (0.13)															
	Lexical density																											
	Task 1	Task 2																										
Band 7	0.82 (0.15)	0.75 (0.15)																										
Band 8	0.82 (0.21)	0.69 (0.13)																										
Syntactic complexity	<p>From test provider: To achieve Band 7, candidates are expected to use a variety of complex structures and produce frequent error-free sentences.</p> <p>From research studies:</p> <ul style="list-style-type: none">• T-unit (Independent clause plus all its associated dependent clauses; minimal terminable unit of text)• Main clause (Clause containing a subject and a predicate which can stand on its own. (predicate = a phrase headed by a TENSED verb))• Dependent clause (Dependent clauses cannot stand on their own, and can only be present when either embedded in or chained to a main clause). <table><tr><td></td><td colspan="2">Mean clause/t-unit</td></tr><tr><td></td><td>Task 1</td><td>Task 2</td></tr><tr><td>Band 7</td><td>1.51</td><td>2.70</td></tr><tr><td>Band 8</td><td>2.07</td><td>2.44</td></tr></table> <table><tr><td></td><td colspan="2">Mean dependent clause/clause</td></tr><tr><td></td><td>Task 1</td><td>Task 2</td></tr><tr><td>Band 7</td><td>0.34</td><td>0.50</td></tr><tr><td>Band 8</td><td>0.61</td><td>0.58</td></tr></table> <p>(Banerjee, Franceschina & Smith, 2007)</p>		Mean clause/t-unit			Task 1	Task 2	Band 7	1.51	2.70	Band 8	2.07	2.44		Mean dependent clause/clause			Task 1	Task 2	Band 7	0.34	0.50	Band 8	0.61	0.58			
	Mean clause/t-unit																											
	Task 1	Task 2																										
Band 7	1.51	2.70																										
Band 8	2.07	2.44																										
	Mean dependent clause/clause																											
	Task 1	Task 2																										
Band 7	0.34	0.50																										
Band 8	0.61	0.58																										
Grammatical accuracy	<p>From test provider: To achieve Band 7, candidates are expected to have good control of grammar and punctuation.</p> <p>From research studies: Target-like use ratio (Ellis and Barkhuizen, 2005)</p> <table><tr><td></td><td>Band 7</td><td>Band 8</td></tr><tr><td>Usage of 'this'</td><td>0.97</td><td>1</td></tr><tr><td>Usage of 'that'</td><td>1</td><td>N/A</td></tr><tr><td>Usage of 'these'</td><td>0.94</td><td>1</td></tr><tr><td>Usage of 'those'</td><td>1</td><td>1</td></tr><tr><td>Copula (am, is, are)</td><td>0.97</td><td>0.86</td></tr><tr><td>Copula (was, were)</td><td>0.97</td><td>N/A</td></tr><tr><td>Subject-Verb agreement</td><td>0.84</td><td>1</td></tr><tr><td>Passives</td><td>0.77</td><td>1</td></tr></table> <p>(Banerjee, Franceschina & Smith, 2007)</p>		Band 7	Band 8	Usage of 'this'	0.97	1	Usage of 'that'	1	N/A	Usage of 'these'	0.94	1	Usage of 'those'	1	1	Copula (am, is, are)	0.97	0.86	Copula (was, were)	0.97	N/A	Subject-Verb agreement	0.84	1	Passives	0.77	1
	Band 7	Band 8																										
Usage of 'this'	0.97	1																										
Usage of 'that'	1	N/A																										
Usage of 'these'	0.94	1																										
Usage of 'those'	1	1																										
Copula (am, is, are)	0.97	0.86																										
Copula (was, were)	0.97	N/A																										
Subject-Verb agreement	0.84	1																										
Passives	0.77	1																										
Coherence and cohesion	<p>From test provider: To achieve Band 7, candidates are expected to logically organise information and ideas; there is clear progression throughout and use a range of cohesive devices appropriately.</p> <p>From research studies: Frequency of use of demonstratives (this, that, these, those)</p> <ul style="list-style-type: none">• Test-takers are more likely to use demonstratives when responding to IELTS Task 2 than to IELTS Task 1.• Use of demonstratives appears to tail off at higher levels of language proficiency, suggesting that other cohesive ties come into use. We would suggest that writers at higher IELTS band levels are more likely to use lexical ties to create cohesion. We would therefore expect performances at higher IELTS band levels to display greater lexical variation and sophistication. <p>(Banerjee, Franceschina & Smith, 2007: 21)</p>																											
Scoring																												
Scoring approach	Analytical																											
Scoring criteria	Candidates are assessed on their performance on each task by certificated IELTS examiners according to the four criteria of the IELTS Writing Test Band Descriptors (task achievement/response, coherence and cohesion, lexical resource, grammatical range and accuracy). Scores are reported in whole and half bands. The public version of the band descriptors can be found at www.ielts.org/researchers/score_processing_and_reporting.aspx																											
Reliability	0.81–0.89 (Figure provided by test providers)																											

Table 1. Context validity parameters (Shaw and Weir, 2007)

<i>Context validity parameters</i>	
Task features	<ul style="list-style-type: none"> • <i>number of tasks</i> • <i>response format</i> • <i>genre</i> • <i>task description</i> • <i>source texts (if any)</i> • <i>domain</i> • <i>topic</i> • <i>communicative purpose</i> • <i>knowledge of criteria:</i> • <i>writer-reader relationship</i> • <i>timing</i> • <i>text length</i> • <i>weighting</i> • <i>delivery mode: pen-and-paper; computer-delivered; online</i>
Linguistic demands of the output	<ul style="list-style-type: none"> • <i>discourse mode</i> • <i>lexical complexity</i> • <i>syntactic complexity</i> • <i>grammatical accuracy</i> • <i>degree of coherence and cohesion</i>
Scoring	<ul style="list-style-type: none"> • <i>scoring approach</i> • <i>scoring criteria</i> • <i>reliability</i>

Table 2. Cognitive validity parameters (Shaw and Weir, 2007)

<i>Cognitive validity parameters: Writing processes</i>
<i>Macro-planning:</i> Gathering of ideas and identification of major constraints (genre, readership, goals).
<i>Organisation:</i> Ordering the ideas; identifying relationships between them; determining which are central to the goals of the written text and which are of secondary importance.
<i>Micro-planning:</i> Focusing on the part of the written text that is about to be produced. Here, the planning takes place on at least two levels: at the level of the paragraph, itself aligned with the overall goal of the writing activity; within the paragraph, the immediate need to structure an upcoming sentence in terms of information.
<i>Translation:</i> Propositional content previously held in abstract form is converted to linguistic form.
<i>Monitoring and revising:</i> At a basic level, monitoring in writing production involves checking the mechanical accuracy of spelling, punctuation and syntax. At a more advanced level, it can involve examining the written text to determine whether it reflects the writer's intentions and fits the developing argument structure of the text. As a result of monitoring, a writer usually returns to those aspects of the text considered unsatisfactory and make corrections or adjustments accordingly.

Table 3. Summary of screening results according to the eight *a priori* criteria

	N=26 ELP tests	
	Yes	No*
1. Does the test have high-level security procedures for test production and delivery?	20	6
2. Does the test have a verification system that allows score users to check and confirm test results?	21	5
3. Does the test assess all 4 communicative skills (reading, writing, listening, speaking)?	26	0
4. Are parallel forms of the test available for all test components?	24	2
5. Is there evidence of the reliability of the test and the dependability of its scores?	19	7
6. Is the approved test centre (or centre network) routinely monitored and inspected for quality assurance?	26	0
7. Is there demonstrable evidence, e.g. publicly available technical or research reports, that the test measures at the CEFR C1 level and/or above?	13	13
8. Is the test domain appropriate (e.g. academic, professional or medical)?	25	1

NB. * no information available on the website; test provider did not reply to or provide sufficient evidence in response to researchers' inquiries.

Table 4. Summary results of test scrutiny in Phase 1 (CPE, CAE, PTE-A, TOEFL-IBT and OET)

Writing Test	CPE	CAE	PTE-A	TOEFL iBT	OET
Overall evaluation	>	=	>	>	<
Features of the Writing Test					
Number of tasks	=	=	>	=	<
Response format and genre	=	=	>	=	<
Source texts	>	>	>	>	>
Domain	=	=	=	=	=
Topic	>	>	=	=	=
Communicative purpose	>	>	=	=	=
Knowledge of criteria	=	=	=	=	=
Writer-reader relationship	=	<	=	=	>
Timing & Text length of expected response	>	<	>	>	<
Skills focus	>	=	>	=	=
Level of cognitive processing	>	=	=	=	<
Linguistic demands of the output					
Discourse mode (rhetorical task)	>	=	=	=	<
Lexical Level	>	=	=	=	=
Syntactic complexity	>	=	=	=	=
Grammatical accuracy	=	=	=	=	=
Coherence and cohesion	>	=	=	=	=
Reliability: Estimate of rating quality (rater consistency)					
Inter-rater reliability* (except for PTE-A which uses automated scoring)	0.73	0.79	0.91	0.74	0.89

KEY: = comparable with IELTS > more demanding than IELTS < less demanding than IELTS

- based on reliability information in the public domain