Identifying the appropriate IELTS score levels for IMG applicants to the GMC register

Dr. Vivien Berry
Professor Barry O'Sullivan
Ms Sandra Rugea

Centre for Language Assessment Research (CLARe)
The University of Roehampton

Report submitted to the General Medical Council

February 2013

Acknowledgements

We are grateful to Cambridge ESOL for providing the examination materials that were used in this study.

Grateful acknowledgements are also due to all the panel members who participated in the study.

Special thanks to those who set up or helped to organise panels.

TABLE OF CONTENTS

Executive summary

Aims

- to address concerns regarding the language competence of non-native English speaking medical practitioners in the UK, this research investigates the current required IELTS levels to determine if they are adequate in light of issues of patient safety
- to investigate the issue of requiring evidence of English language ability from all non-native English speaking medical practitioners seeking admission to the GMC register

Objectives

- to determine if the current overall IELTS score of Band 7, with no separate skill score lower than Band 7, is adequate as a preliminary language screening device for International Medical Graduates (IMGs)
- to determine if European Economic Areas graduates (EEAs) should provide the same evidence as IMGs if evidence of English language competence should ever be required in order for them to be admitted to the GMC register
- to determine if the IELTS test, by and of itself, provides an adequate measure of English language ability for overseas medical practitioners seeking admission to the GMC register

Methods

1. A literature review was conducted to
    - identify current practice in language assessment for migrant medical professionals across Europe and in international countries where English is a first or joint official language
    - locate and review any existing research that has been undertaken on the language needs of these medical professionals, both in the UK and internationally
    - identify current and historical thinking in the area of standard setting as an empirical methodology
    - identify the most appropriate standard setting approach for this project and to identify sources to support a decision to use this approach
2. Eleven initial stakeholder panels were convened comprising doctors x 3 (15 participants); nurses x 3 (15 participants); public/patients x 3 (20 participants); Allied Health Professionals x 1 (5 participants); Responsible Officers/Medical Directors x 1 (7 participants). A total of 62 people, 30 males and 32 females, 53 white and 9 ethnic minorities, ranging in age from early 20s to late 60s, from various regions of the UK, participated.

3. Findings from the initial panels were analysed and presented to a final panel for confirmation and development of recommendations. The final panel comprised 2 doctors, 2 nurses, 2 patients, 1 AHP and 1 RO/MD.

Findings

1. The current overall IELTS score of Band 7, with no separate skill score lower than Band 7, is not adequate as a preliminary language screening device for International Medical Graduates (IMGs)
2. If evidence of English language competence should ever be required in order for EEAs to be admitted to the GMC register, they should provide the same level of evidence as IMGs
3. The IELTS test provides an adequate measure of English language ability for overseas medical practitioners seeking admission to the GMC register,

Recommendations

Despite the criticisms made of the IELTS test, which are mainly concerned with the appropriacy of using a test designed for use in the academic domain in other domains (McNamara, 2000; McNamara and Roever, 2006), IELTS has been shown to be a robust general language proficiency test. This is reinforced by comments from the initial stakeholders' panels and from the Final Panel, most of whom have criticised certain aspects of the test but most of whom have also stated that it is a very good test of general English ability. It is clear, therefore, that although obviously IELTS cannot be used to assess medical competence, it is certainly an appropriate instrument to use to assess the language ability of prospective medical doctors to the GMC register.

RECOMMENDATION 1: In light of the demonstrated need for medical professionals to have a high level of general English language ability in addition to accepted medical qualifications, we recommend that the IELTS test should be retained as an appropriate test of the English language competence of overseas-trained doctors.

The Final Panel, in keeping with several of the initial stakeholder panels, found it very difficult to arrive at an overall band score to recommend, many participants preferring to specify a profile. This is supported by research findings such as those of Chalhoub-Deville and Turner (2000). However, as an overall band score is an essential component of IELTS reporting, it was agreed that an average of the profile scores would be recommended, which is how the overall IELTS score is computed.

RECOMMENDATION 2: We recommend that the band score level requirements for IELTS should be revised and that the GMC should consider adopting the following profile which reflects the importance of oral skills, with listening being of paramount importance, but allows for some flexibility in assessing written skills:

Overall	Band 8
Listening	Band 8.5
Speaking	Band 8
Reading	Band 7.5
Writing	Band 7.5

Despite an EU directive which currently exempts graduates with acceptable medical qualifications from within the European Economic Area and Switzerland from providing evidence of English language ability when applying for registration in the UK, this is not considered an acceptable situation with regard to patient safety. It is argued that both IMGs and EEA graduates should provide evidence of English language ability in order to register with the GMC.

RECOMMENDATION 3: We recommend that the GMC should attempt to find a way of requiring all non-native speakers of English to provide evidence of English language competence before being allowed to practise medicine in the UK and that if ever it becomes possible to require evidence of language ability from EEA graduates, they should provide the same evidence as IMGs.

Suggestions for further consideration

In terms of IMG candidates applying for admission to the GMC register, it may be that language is not the only issue and perhaps there should be a closer look at the relationship between language ability as assessed by IELTS and medical knowledge as assessed by PLAB.

SUGGESTION 1: In order to determine whether or not there is a strong relationship between performance on the IELTS test and success in each part of the PLAB test, we suggest that the GMC consider statistically analysing the relationship between score performance on IELTS and score performance on PLAB Parts 1 and 2.

It may also be useful to consider the argument that the IELTS test, particularly the listening test, does not provide adequate evidence of the skills previously identified as being essential for a doctor. Discussing the Canadian medical context, Watt et al. (2003:35) recommend that Canada works to develop and implement "a national language standard and standard assessment procedures for demonstrating the language proficiency required for medical practice in Canada" and suggest that the language standard be one that "parallels the existing Canadian Language Benchmarks (CLB), and is adjusted to more accurately reflect the language demands of working in medical contexts". There is no reason why the same could not also be done for the UK as the UK Occupational Language Standards already exist and could readily be adapted and extended to cater for medical practice.

SUGGESTION 2: We suggest that the GMC considers developing a UK language standard and standard assessment procedures for demonstrating the language proficiency required for medical practice in the UK.

1. Introduction

1.1. Background

Most non-native speakers of English who hold medical qualifications obtained in countries other than the UK or those of the European Economic Area and Switzerland, known as International Medical Graduates (IMGs), provide evidence of English language ability by achieving a minimum overall score, currently set at Band 7, on the academic module of the IELTS (International English Language Testing System) test in a single sitting, with no separate skill score lower than Band 7. Having achieved this, they are then eligible to register to take Part 1 of the Professional and Linguistics Assessment Board (PLAB) test which must be passed in order for them to be eligible to take Part 2 of the PLAB test, success in which makes them eligible for registration with the GMC to practice medicine in the UK. Although there are a number of alternative routes to registration[1], the majority of IMG applicants follow the IELTS – PLAB route to registration with the GMC.

At present, in accordance with the Medical Act, 1983, graduates with acceptable medical qualifications from within the European Economic Area and Switzerland (EEAs) are exempt from providing evidence of English language ability when applying for registration in the UK. However, in 2010, in response to the European Commission's evaluation of Directive 2005/36/EC[2] on the mutual recognition of professional qualifications, 28 competent authorities from 23 member states created an informal network to stimulate discussions and support the drafting of national experience reports on the Directive. Further to their meetings and the exchange of experiences in relation to the evaluation of the Directive, on 13 September 2010 they collectively issued the Berlin Statement[3] which called on the Commission to (*inter alia*):

> "Examine the language provisions in the Directive to address the concerns of competent authorities in relation to language proficiency of migrant doctors in the interest of patient safety."

Since the well-publicized incident in the UK in 2008 when a German-registered doctor with questionable English language ability provided less than adequate locum care to a patient who subsequently died, there has been much public discussion about the different language requirements for IMGs and EEAs. As neither category of potential registrants are native English speakers, the issue of language ability perhaps ought to be considered

---

[1] See http://www.gmc-uk.org/doctors/registration_applications/routeG.asp

[2] Directive 2005/36/EC – the recognition of professional qualifications sets out the UK obligations for recognising the professional qualifications held by professionals from within the EEA. The Directive is currently under review by the European Commission and may change the recognition requirements for EEA nationals. Negotiations between the Commission, Member States, and the European Parliament commenced in December 2011. The Commission anticipates that a new Directive will be agreed by the end of 2012 but is unlikely to be implemented in the UK before 2014 (two years after its adoption by the EU Institutions. (Taken from 'Language_competency_200212.pdf', published by NHS Employers).

[3] see http://ec.europa.eu/internal_market/qualifications/docs/evaluation/experience-report-doctor_en.pdf for the Berlin Statement and experience reports from national authorities with regard to doctors

as a “fitness to practise” issue rather than simply coming under a directive of nationality and acquired rights (EEA or not).

1. 2. Aims and objectives of the research

In order to address the concerns outlined above regarding the language competence of non-native English speaking medical practitioners in the UK and, in addition, to confirm, strengthen and extend the findings of a previous study (Banerjee, 2004), this research investigates the current required IELTS levels to determine if they are adequate in light of issues of patient safety; we also investigate the issue of requiring evidence of English language ability from all non-native English speaking medical practitioners seeking admission to the GMC register.

1.3. Research Team

The project has been delivered by a team of researchers who understand the complexity of this type of project, having worked with examination boards in the UK and overseas. The researchers have broad experience of working on projects with clients from different cultures and different disciplines. They have acted as consultants to Cambridge ESOL on assessment matters over a period of five years, advising on a range of examinations including IELTS; they have also successfully completed a number of IELTS Joint Funded Research projects and standard-setting projects, so are very familiar with the area of focus required for this study.

1.4. Research Approach

The detailed approach to the standard-setting aspect of this project was designed to make the entire process both transparent and theoretically sound. The use of quantitative probability-based statistical analysis in tandem with a detailed qualitative analysis of the data gives rise to recommendations that can be shown to be valid and robust. In addition to the innovative use of many-faceted Rasch (MFR) statistical analysis to support the decisions of the stakeholder focus groups (SFGs) the approach included an additional validation process, whereby the recommendations were debated by an expert focus group (EFG) who made the final recommendations included in this report. Figure A shows the basic design of the approach.

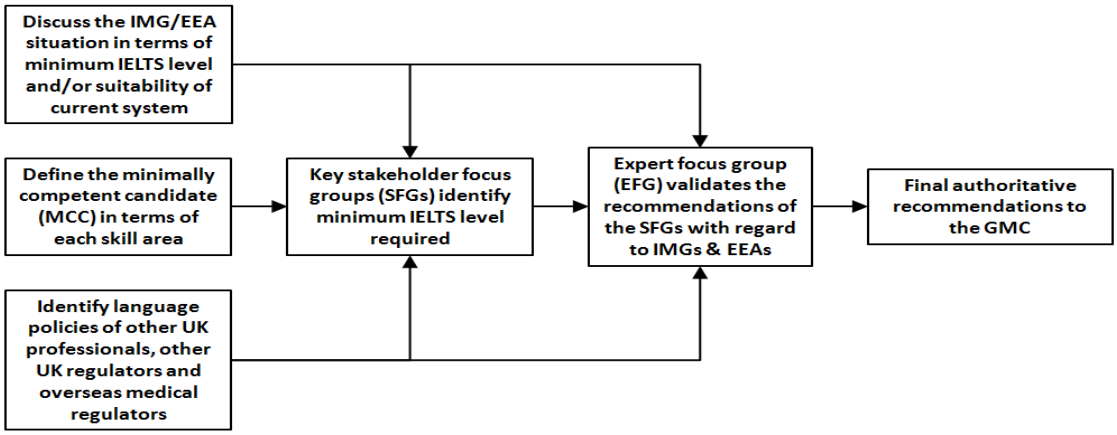


Figure A Approach

1.5. Ethical considerations

1.5.1. NHS Research Ethics Committee

In accordance with the regulations of the NHS Research Ethics Service (NRES), this project is considered to be a Service Evaluation rather than Research as defined in the NRES publication *Defining Research* 4. In addition, no patients were recruited through the auspices of the NHS. As such, ethics approval was not required from the NHS Research Ethics Committee.

1.5.2. University of Roehampton Research Ethics Committee

As the project was undertaken by members of CLARe, ethical approval was granted by the University of Roehampton's Research Ethics Committee. Questions addressed in the application included, *inter alia*, presenting full details of:

- Method of recruitment of participants
- Preservation of participants' anonymity
- Participants' informed consent
- Risks and benefits to participants
- Storage of data
- How ethical considerations arising from the project will be handled
- Dissemination of findings

1.5.3. Replication of previous research

This project may possibly be thought of as a replication of a project originally undertaken in 2004 on behalf of the GMC by Dr. Jayanti Banerjee from Lancaster University with the assistance of Dr. Lynda Taylor from Cambridge ESOL. However there are numerous differences which include:

- the 2004 study had seventeen participants on three panels; the 2012 study had sixty-two participants on eleven separate panels
- the 2004 study was carried out in one location; the 2012 study was carried out in seven geographically diverse locations covering each country of the UK
- the 2004 study focused on only the two productive skills, writing and speaking; the 2012 study considered all four skills, the receptive skills of reading and listening as well as the productive skills of writing and speaking
- the 2004 study was only concerned with advising on appropriate IELTS levels for different categories of IMG applicants to the GMC register; the 2012 study considered both IMG and EEA applicants
- the 2004 study accepted at face value the integrity of IELTS as an assessment instrument for professional purposes; the 2012 study investigated whether IELTS is, in fact, considered to be an appropriate instrument to assess the language ability of medical professionals and made recommendations based on qualitative analysis of the data obtained from panel discussions
- to the best of the current researchers' knowledge, findings from the 2004 study have not been disseminated publicly with the exception of one 20-minute paper

4 See http://www.nres.npsa.nhs.uk/applications/is-your-project-research/

presentation at the Language Testing Research Colloquium (LTRC) in Ottawa, Canada in 2005 (Banerjee and Taylor, 2005); it was not therefore thought desirable to attempt to replicate the original study due to its limited focus and small number of judges, as outlined above.

2. Literature review

2.1. Aims of the literature review

The review looks at the published literature in the areas of medical language assessment and the broader area of test standard-setting, of which this study is seen as a variant. Our approach has therefore been two-fold although the overall aims and objectives for each are identical, namely, to provide a thorough, comprehensive and up-to-date review of the major published findings in each of the two aspects.

The specific aims of these two aspects are:

1. To identify current practice in language assessment for migrant medical professionals across Europe and in international countries where English is a first or joint official language.
2. To locate and review any existing research that has been undertaken on the language needs of these medical professionals, both in the UK and internationally.
3. To identify current and historical thinking in the area of standard setting as an empirical methodology.
4. To identify the most appropriate standard setting approach for this project and to identify sources to support a decision to use this approach.

2.2. Search strategies

The search strategies used for both the medical language aspects and standard setting of the literature review included:

- Print literature including book chapters, peer-reviewed journal articles, published research reports
- Internet search engines including British Education Index, EduSearch, ERIC, Google Scholar, JStor (for archived work), Medscape, NCBI, SearchEdu and PubMed. Search terms included:

Med*	Conversation*
Practitioner*	Language*
Doctor*	Nurse*
Discourse*	Needs*
Communication*	Analys*
Interact*	Migra*

For each of the print and internet searches, the only exclusion criteria applied were those of relevance and reputation. Literature published in a language other than English was also excluded from the internet search although we attempted to deal with this last criterion through an alternative strategy.

In terms of relevance and reputation, non-peer reviewed print literature was excluded unless it emanated from a well-regarded source and was authored by respected researchers in the field. Examples of reports included in this category are research reports published by Cambridge ESOL following funded research projects investigating aspects of IELTS (the initial application for funding of which is peer-reviewed) and the

experience reports from competent national authorities with regard to doctors. Age of publication was not, in itself, considered to be a criterion for exclusion as much of the recent work done in both standard setting and medical language has developed from seminal work first published many years ago.

The literature search on the language requirements of migrant doctors focused on finding any reported work on the identification or analysis of the language needs of medical practitioners, specifically in the UK and Europe, but also in the USA, Canada, Australia and New Zealand where there has been much work done in the area of medical discourse analysis.

Additional resources used in the literature search included:

- the considerable experience of the main researchers and their graduate students
- suggestions from colleagues who also have extensive experience in either standard setting, medical English, or both
- contact with colleagues in EU countries to elicit regional knowledge, in particular information published in local languages which would not have been thrown up in the internet search
- review of language practices of medical registration councils in Anglophone countries with particular reference to the IELTS (or equivalent) levels required for IMGs
- review of language policies of other major registration organizations in UK
- a survey of local language practices of medical registration bodies for IMGs and EEAs in non-English speaking EU countries

2.3. Medical language, communication and assessment

The importance of effective communication between doctors and patients and others involved in the professional medical workplace has been recognized through the recent introduction of communication skills training into the medical curriculum in the United Kingdom (von Fragstein et al., 2008). Research on the language needs of medical professionals has highlighted the critical nature of communication in the medical profession, see for example the work on the link between language barriers and critical events in medical care (Baker and Robson, 2012; Flores et al., 2002; Flores et al., 2003; Cohen et al., 2005; Roberts et al., 2005; Johnstone & Kanitsaki, 2006; Divi et al., 2007). The often crucial, and little understood, role in facilitating communication played by interpreters in medical interactions has been highlighted by researchers such as Davidson (2000) and Ku & Flores (2005). The importance of reciprocity of doctor-patient communication in terms of social differences in the physician-patient relationship is addressed in Verlinde et al. (2012) .

Issues in the education of medical professionals have been highlighted by Candlin & Candlin (2003) and addressed in studies by Wood & Head (2004) and Drew et al. (2001) who support the argument for an empirical basis for such training; using tasks which

replicate the types of interactions which typify the domain. Other studies which have implications for medical communication education include those of Hoekje (2007), McNeilis (2001), Silverman et al. (2005), Skelton et al. (2001), and Skelton (2008), all of whom support the argument for using data garnered from medical communication events to support the learning process. A recent empirical study by Elder et al. (2012) confirms that appropriate language skills are absolutely fundamental to medical trainees' performance in face-to-face interactions with patients, and notes that patient-centred consultations are completed much more effectively when trainees are able to appropriately use, for example, lay rather than medical language. A forthcoming paper comparing methods of assessing the English language proficiency of health professionals, mainly doctors, seeking to practice in the UK, North America and Australia, concludes that the inter-relationship between language proficiency, communicative competence and clinical communication skills is an extremely complex problem which requires much further study (Taylor and Pill, forthcoming, 2014).

2.3.1. Criticisms of language assessment for the medical domain

The earliest criticisms of how the assessment of language for health professionals was carried out in the USA and the UK came from researchers working in the health professions. This criticism, of the use of what were perceived as international general purpose proficiency examinations, has continued, though its source has spread to include researchers in the area of applied linguistics.

In the early 1990s, Freidman et al. (1991) examined both the spoken proficiency and clinical competence of overseas medical graduates and found that the language examination used (TOEFL) offered an insufficiently complete estimate of the English ability of candidates in specific medical contexts. This criticism was echoed by both Chur-Hansen (1997) and Whelan et al. (2001), who acknowledged that although what they see as international general-purpose proficiency examinations (TOEFL and IELTS – actually both are intended to assess academic language) may function well as predictors of general language proficiency, they are of little value in predicting ability in medical communication. When exploring the self-expressed language needs of trainee medical professionals, Lepetit and Cichocki (2002) report that that oral skills were considered the most relevant to the medical professionals in their study, a finding also reflected in the work of Lear (2003), working in this case with Spanish rather than English (the focus of the other studies reported on here). This finding adds to the debate by highlighting the need to move away from a conception of overall level (e.g. Band 7 on IELTS; C1, a level of the CEFR; or 237, a TOEFL score) to the realisation that different language sub-skills may prove to be important for individuals engaged in different professional roles.

The use of the IELTS for the purpose of high-stakes decision-making in the medical language domain has been criticised in the language testing literature by McNamara (2000), McNamara & Roever (2006), Reed and Wette (2009) and Wette (2011). The main criticism lies in the fact that the IELTS was originally designed for use in the academic domain and is not considered appropriate for use in other domains without substantial empirically-derived evidence in support of this usage.

Early published research into the development of a test designed to evaluate both the professional and the language abilities of overseas doctors is that of Rea-Dickens (1987)

who describes the development of the Temporary Registration Assessment Board (TRAB), the forerunner to the PLAB. The development of TRAB was innovative at the time as linguists and medical experts worked collaboratively to analyze the language used by health professionals in British hospitals and used this knowledge to inform decisions on test content. Another study published at around the same time, also emphasizing the importance of close collaboration between linguists and health professionals, was that of Alderson et al. (1986) in relation to the revision of the Australian Occupational English Test (OET). Much later, in the Canadian context, Watt et al. (2003) suggest that the tasks included in tests for medical professionals should reflect the physical setting and performance parameters (e.g. location, audience and other characteristics of task performance) of tasks performed in the real world of medical practice, referred to in the literature as 'situational authenticity'. They also felt that such test tasks should mirror the communication or interaction type found in these medical communication settings and that the procedures for assessing performance (who should perform the assessing, what aspects of language the rating scale/rubric should focus on) should also reflect the real world of medical practice. Lear (op. cit.) also implies that this need for situational authenticity should be reflected in medical language assessment. However, if language tests are developed according to the idealised criteria of situational authenticity, as seems to be the consensus of recent research , a problem then arises with assessment and with the fact that assessors may need to be both language aware and highly content-familiar (Harding et al. 2011, Pill and Woodward-Kron, 2012).

In addition, there has been some concern with the need for valid and empirically derived standards of language ability in specific medical domains (Shapiro et al. 1989; Haertel, 1999; Jacoby & McNamara, 1999; Whelan et al., 2001; O'Neil et al., 2005; O'Neil et al., 2007), which includes research on how the language is used in the domain (Ali, 2003; Wodak, 2005; Heritage & Maynard, 2006). The weakness of much of this work, with the possible exception of O'Neil et al. (2007), has been the failure to translate the research findings into operational procedures. This has had the effect of significantly limiting the impact and value of much research and increasing the perception in the general population of the academy as an 'ivory tower' where individuals debate issues of little or no 'real world' import.

2.3.2. Questions arising from this part of the review

Three questions arise out of this:

1. Does the IELTS offer an appropriate measure of the language ability of prospective medical professionals?
2. If it is found that IELTS is an appropriate instrument, what is the most appropriate level for prospective medical doctors to attain before being accepted to practice in the UK?
3. Should all non-native-English speaking doctors be asked to empirically demonstrate their language level?

When we consider how medical professionals are assessed using international proficiency tests, the following quote (O'Sullivan, 2012: 75) highlights both the disquiet of some members of the language assessment academy and their failure to date to respond to their own concerns:

> *There has been a disquieting lack of empirical evidence to validate the decision to use these tests for uses other than those for which they have been designed (in the case of IELTS, academic English). On the other hand, it should be stressed that no empirical evidence has been published to cast doubt on the validity of existing language tests for immigration.*

This study aims to both respond to the first two of these questions and to begin the process of building a substantive validity argument (Messick, 1989) in order to either support or reject the use of the IELTS test and, in addition, to set acceptable IELTS band levels where this is seen as appropriate. The study also addresses for the first time the question of the language skills of all medical doctors practising in Britain.

2.4. Standard setting

Standard-setting has been defined by Cizek (1993: 100) 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 states or degrees of performance*

In other words, standard setting is the process by which we establish the critical boundary point or points which will be used to interpret test performance. An example of this might be the setting of a pass/fail boundary for a mathematics examination, establishing empirical evidence of a link between a pass/fail boundary on an educational examination and a set of specific learning standards (e.g. see O'Sullivan, 2009; Kantarcıoğlu et al., 2010, Kantarcıoğlu, 2012), or, in this case, the criterion level on a particular test (the IELTS test) above which medical doctors must perform in order to be allowed to practise medicine in Britain.

2.4.1. Approaches to standard-setting

In her exhaustive overview of the area, Kaftandjieva (2004) reports that the most commonly used approaches to standard-setting are test-centred. In these approaches the focus is on the test, as opposed to the test takers. They involve a number of judges who are asked to review a test or a section of a test and to estimate the likely difficulty of each item (or question). Advantages to the approaches include their relative ease of application, the fact that they can involve a large number of judges (thus increasing the reliability of the resultant decision) and the fact that they can, in general, be applied when no empirical data (which could be used, for example, to demonstrate to judges the impact of their decision) is available.

The approaches have been criticised mainly with regard to the difficulty experienced by judges in accurately and consistently estimating item difficulty, possibly due to differing interpretations of the questions under review and/or the criterion level to be applied (see for example Impara & Plake, 1998; Chang, 1999; Plake & Impara, 2001).

Kaftandjieva (2004: 15) suggests a number of ways of dealing with this. These can be summarised as:

- Extensive appropriate training (though she does not indicate what this might entail)

- A validity check (again unspecified, though given her general approach, this is most likely to be statistically driven)
- Some adjustment to empirical data should be made (again likely to be statistically based)
- Employ more than one method
- Include non-statistical data or information when reaching a final decision

The other approach to standard-setting is examinee-centred. As the name implies, this approach uses as its basis examples of actual examinee/candidate behaviour, and in a test of speaking this would include either audio recordings of spoken language or copies of responses to writing tasks. The main advantage of methods that employ such an approach is that it is far easier for judges to evaluate concrete examples of performance than to estimate the possible difficulty of a test question (a far more abstract process).

Since the approach taken will depend to a large degree on the type of test (for example the test-centred approach is more appropriate to tests of receptive language skills, while examinee-centred approaches are more appropriate to productive skills tests) it is clear that a number of different approaches will be required during the current project.

The other aspects of standard-setting which are important to any project involving such an approach to critical boundary setting are related to establishing the number of judges to use and an approach to ensuring the accuracy and consistency of the judgments made. With regard to the former, it is widely accepted that the larger the number of judges the more likely will be a valid and reliable outcome (Kane, 1994: 439). The actual number of judges is a matter of some debate, ranging from not less than 5 (Livingston & Zieky, 1982) to approximately 7 to 11 (Maurer, et al., 1991; Biddle, 1993) to as many as 15 (Hurtz & Hertz, 1999). Papagiorgiou (2007), cautions against larger groups as his evidence suggests that in such groups the impact of group dynamics may prove to be destabilising as some participants may tend to dominate the judgment process. This suggests that a number of groups working independently with a single, final decision-making group is the optimum.

Kozaki (2004), O'Sullivan (2009) and Papagiorgiou (2007) all explored the use of probability-based statistical procedures (many-facet Rasch, also known as multi-faceted Rasch, – both using the acronym MFR) to help establish the cut scores in language tests. In the case of Kozaki (2004), who was attempting to set cut scores for an assessment of medical translators, a process called generalizability-theory (G-theory) was also used. The use of MFR deals well with the issue raised by Kaftandjieva (2004: 15) with regard to adjustment of the overall judgment made by a group of judges by taking into account a number of variables. The resultant cut score or decision point, is therefore more likely to reflect the true underlying level intended by the group.

2.4.2. Identifying the appropriate standard-setting approaches

Because the IELTS test comprises two receptive skills papers (listening and reading) and two productive skills papers (speaking and writing), two approaches to standard-setting are required.

The first of these approaches is that applied to the receptive papers. The most commonly used approaches include

- The modified Angoff method (Hambleton & Plake, 1995).
- Yes/No method, which is a variant of the Angoff method and modified Angoff methods (Impara and Plake, 1998)

Since neither of these methods result in exact cut-scores (i.e. they are not integers), the results need to be rounded to the nearest integer. Cizek and Bunch (2007) suggest two ways of doing this. The first involves rounding to the nearest score (up or down), while the second one means rounding up to the next integer. The latter approach represents the more conservative option and is likely to be more appropriate to this project, where it is important to be certain that only really suitable candidates are deemed acceptable. It would certainly be consistent with the current scoring of Part 2 of the PLAB test where one standard error of measurement (SEM) is added to the cut score to reduce false positives in the interests of patient safety and which a recent report recommends as appropriate (McLachlan et al., 2012: 56-57).

For a fuller outline of the Modified Angoff approach see Cizek & Bunch (2007: 83), while the Yes/No method is also discussed by Cizek & Bunch (ibid: 88-89).

When it comes to standard-setting for performance-based language tests (i.e. writing and speaking), the most appropriate method will be an examinee-centred approach. With regard to the requirements of this project, the most appropriate of these appears to be the 'Examinee Paper Selection' method (Hambleton, et al., 2000), also known as the Benchmark method (Faggen, 1994). In this approach, judges are asked to make decisions based on test task performance (i.e. recordings of speech or copies of written work) rather than by reviewing the task input material from the test paper. Typically, the group of panel members will first operationally define the person who just meets the required standard (referred to in the educational measurement literature as the 'minimally acceptable person' and in the language assessment literature as the 'minimally competent candidate'). Panel members typically make judgments on a set of test performances and then discuss their decisions before progressing to another round of judgments. The process is repeated until an agreed decision is arrived at.

2.4.3. Ensuring consistency and agreement of judges

Consistency and agreement among participants is vital in a project in which a critical boundary is to be set based primarily on subjective judgment. While the traditional approach to establishing evidence of this has tended to focus only on the first element (consistency), it is equally important that some measure of agreement is also established. A relatively recent approach has been to use a statistical package entitled FACETS (Linacre, 1989). This uses the MFR measurement model (itself an extension of the Rasch model for dichotomous data) to take into account the impact on the probable true value of a judgment of a range of facets (or factors) for which the researcher has data. MFR analysis allows for information to be obtained from each facet included in the analysis and for these output data to be compared. The practical advantage, therefore, is that the programme takes into account a range of factors when proposing a 'fair average' estimate of a set of judgments based on a range of input data.

As Lumley (2005: 95-96) explains:

> *Multi-faceted Rasch measurement proposes a straightforward mathematical relationship, expressed in probabilistic terms, between item or task difficulty, candidate ability, and other relevant elements or facets of the assessment context, such as rater harshness. This relationship maybe expressed as follows:*
>
> *P = B - D - J - K - O (etc.)*
>
> *where*
>
> *P = probability of a given score on rating scale*
>
> *B = ability of the candidate*
>
> *D = difficulty of the task*
>
> *J = harshness of the judge (rater)*
>
> *K = difficulty of a particular level on the rating scale*
>
> *O = other facet of the assessment context*
>
> *All facets are placed on the same scale and described in terms of equal interval units called logits.*
>
> *The fact that Rasch analysis models harshness of each rater as well as difficulty of each task and item, and uses the derived estimates in calculating candidate ability, means that this form of analysis was considered to provide the fairest model of calculating ability...*

Lumley's (2005) explanation of MFR can be adapted to reflect the approach taken in this project, namely:

> Multi-faceted Rasch measurement proposes a straightforward mathematical relationship, expressed in probabilistic terms, between task difficulty, candidate ability, and judge harshness. This relationship maybe expressed as follows:
>
> P = B - D - J - K
>
> where
>
> P = probability of a given judgment on the acceptability scale
>
> B = ability of the candidate
>
> J = harshness of the judge
>
> K = difficulty of a particular level on the acceptability scale
>
> All facets are placed on the same scale and described in terms of equal interval units called logits.
>
> The fact that Rasch analysis models harshness of each judge as well as difficulty of each task reviewed, and uses the derived estimates in calculating candidate acceptability, means that this form of analysis is considered to provide the fairest model of calculating acceptability...

MFR has been used in a variety of language assessment related studies; see McNamara & Knoch (2012) for a useful review of these studies. MFR has also been used in other studies in which expert panel members were asked to contribute to a standard-setting event (see for example Papagiorgiou, 2007; O'Sullivan, 2009; Kantarcıoğlu 2012). For a comprehensive explanation of how MFR works in language testing contexts, see Eckes, 2009.

2.4.4. Issues arising from this part of the review

Three issues arise out of this part of the literature review:

1. The approach to standard-setting should reflect the format of the test papers under review.
2. The number of judgments which contribute to the final recommendation should be large enough to ensure a meaningful decision, be organised in such a way as to limit any group dynamics effect, and be representative of the stakeholder population.
3. The technical aspects of the judgment processes (i.e. accuracy and consistency) are central to the validity of the final judgments made.

2.4.5. Caveat

Concluding her review with some words of caution, Kaftandjieva (2004: 31) points out that:

> *there is no true cut-off score, there is no best standard setting method, there is no perfect training, there is no flawless implementation of any standard setting method on any occasion and there is never sufficiently strong validity evidence.*

In the case of the research study reported here, the establishing of an appropriate cut score for prospective medical practitioners planning to work in Britain is an undertaking with significant social and professional consequences. As this involves the interpretation and use of a test score, it should be seen as an important part of the validation process (where evidence for the interpretation of a test score for a specific purpose is gathered). Since the responsibility for gathering such evidence ultimately lies with the user (Cizek, 2011; O'Sullivan & Weir, 2011), this research will endeavour to respond to the issues and questions raised in this review on behalf of the General Medical Council.

3. Methodology

3.1. Materials used

All IELTS test materials used were supplied by Cambridge ESOL and were either live (currently in use), or standardized for training purposes, IELTS tests.

3.1.1. Writing paper

The IELTS writing paper consists of two tasks. In Task 1, candidates are presented with a graph, table, chart or diagram and are asked to describe, summarise or explain the information in their own words; in Task 2, candidates are asked to write an essay in response to a point of view, argument or problem[5].

Candidates are assessed on their performance on each task by trained 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 (http://www.ielts.org/pdf/Writing%20Band%20descriptors%20Task%202.pdf). Scores for each of the four criteria are equally weighted and are reported in whole and half bands. Half bands are awarded when a candidate receives two of the four scores matching the descriptor of the Band awarded and two scores matching the Band above.

In the 2004 study (Banerjee, 2004; Banerjee and Taylor, 2005), exemplar writing performances were used from Task 1. While both tasks contribute to the overall score awarded for writing, the second task is more substantial (in terms of length) and is double weighted, meaning that it contributes twice as much to the final writing score as Task 1. For this reason, the exemplar performances in this study were represented using Task 2 only. This was done to limit the amount of work for the judges while still gaining as true a picture of the level as possible.

Twelve exemplar writing papers, standardized by Cambridge ESOL and ranging in assessed ability from Band 5 to Band 8.5 were provided. Each paper had been written in response to Task 2 of the IELTS writing test and addressed one of two different prompts.

3.1.2. Speaking paper

The IELTS speaking paper consists of an approximately 15 minute one-to-one interview in three parts: Part 1 is a question-and-answer warm-up phase in which basic information about the candidate is elicited and a few general questions are asked on familiar topics such as home, family, work, studies and interests; Part 2 consists of a 1 - 2 minute monologue on a topic selected by the assessor (with the candidate offered suggestions on what to include and given one minute to prepare) after which the assessor asks 2 - 3 questions related to the topic; Part 3 consists of an extended question and answer section in which the examiner asks further probing questions connected to the topic of Part 2, thus giving the candidate an opportunity to discuss more abstract issues and ideas.

[5] see http://www.ielts.org/pdf/Information_for_Candidates_booklet.pdf for further information about the format of the IELTS test and the various components of it from which the descriptions of the various parts of the test have been taken.

Candidates are assessed on their performance throughout the test by trained IELTS examiners according to the four criteria of the IELTS Speaking Test Band Descriptors: fluency and coherence, lexical resource, grammatical range and accuracy, pronunciation (http://www.ielts.org/pdf/Speaking%20Band%20descriptors.pdf). As with the writing test, scores for each of the four criteria are equally weighted and are reported in whole and half bands. In keeping with the reporting of writing scores, half bands are awarded when a candidate receives two of the four scores matching the descriptor of the Band awarded and two scores matching the Band above.

Sixteen speech samples, standardized by Cambridge ESOL for training purposes and ranging in assessed ability from Band 5 to Band 8.5 were provided.

3.1.3 Reading paper

The IELTS reading paper consists of 3 sections with a total text length of 2,150-2,750 words. Each section contains one long text. Texts are authentic and are taken from books, journals, magazines and newspapers. They have been written for a non-specialist audience and are on academic topics of general interest. According to information provided for prospective candidates, texts are appropriate to, and accessible to, candidates entering undergraduate or postgraduate courses or seeking professional registration (see http://www.ielts.org/pdf/Information_for_Candidates_booklet.pdf). Texts range from the descriptive and factual to the discursive and analytical and may contain non-verbal materials such as diagrams, graphs or illustrations; if texts contain technical terms, a simple glossary is provided.

Each reading test has 40 questions. Many different question types are used, chosen from the following: multiple choice, matching, plan/map/diagram labelling, form completion, note completion, table completion, flow-chart completion, summary completion, sentence completion, short-answer questions. Each correct answer is awarded 1 mark. Scores out of 40 are converted to the IELTS 9-band scale and reported in whole and half bands.

Two full length reading tests were provided by Cambridge ESOL, complete with item statistics (facility value and item discrimination) and raw score to band conversion tables.

3.1.4 Listening paper

The IELTS listening paper consists of 4 sections: Section 1 is a conversation between two people set in an everyday social context (e.g. a conversation in an accommodation agency); Section 2 is a monologue set in an everyday social context (e.g. a speech about local facilities or a talk about the arrangements for meals during a conference); Section 3 is a conversation between up to four people set in an educational or training context (e.g. a university tutor and a student discussing an assignment, or a group of students planning a research project); Section 4 is a monologue on an academic subject (e.g. a university lecture). According to information provided for candidates, a variety of voices and native-speaker accents are used and each section is heard once only (see reference, above, for reading).

Each listening test has 40 questions. A variety of question types is used, chosen from the following: multiple choice, matching, plan/map/diagram labelling, form completion, note

completion, table completion, flow-chart completion, summary completion, sentence completion, short-answer questions. Each correct answer is awarded 1 mark. Scores out of 40 are converted to the IELTS 9-band scale and reported in whole and half bands.

Two full length listening tests were provided by Cambridge ESOL, complete with item statistics (facility value and item discrimination) and raw score to band conversion tables.

3.1.5. Additional materials

IELTS band descriptors for writing and speaking skills + CEFR[6] 'can do' statements for C1 and C2 level reading, writing, listening and speaking were adapted to supplement the 'can do' statements produced by the panels (see Appendix 1 for 'can do' statements for each skill).

3.2. Initial Stakeholder panels

3.2.1. Recruitment of participants to initial stakeholder panels

Recruitment of individual panels was effected in several different ways: the doctors' panel in Scotland was recruited by an individual doctor who responded to a GMC request for assistance; the ROs/Medical Directors' panel in Birmingham was recruited with the help of GMC contacts in the West Midlands; the nurses' panel in Northern Ireland was recruited with the assistance of the Royal College of Nursing, Northern Ireland. With the exception of issuing the initial request for assistance and providing a list of contacts for ROs and Medical Directors, the GMC had no further involvement in recruitment. All other doctors', nurses', patients'/public and allied health professionals' panels were recruited through the personal contacts of the researchers. It must be emphasized, however, that only one contact for each panel was known to the researchers; this individual recruited the other members of their panel, all of whom were entirely unknown to the researchers. With the exception of the ROs/MDs who were offered hospitality, hotel and travelling expenses only, all other panel members were paid an incentive of £310 per day for their participation.

3.2.2. Sampling approach

In each of the initial stakeholder panels we aimed to achieve the following characteristics:

- Mixed age range
- Both male and female participants
- Ethnic diversity
- Both urban and rural inhabitants

In addition, for the doctors' panels we aimed to achieve the following:

- Mixed level of seniority, Foundation Year 1-2 up to ST 7/8 including consultants

[6] CEFR is the Council of Europe Framework of Reference for Languages

- Mixed settings including:
  Primary care
  Hospitals
  Supporting healthcare services e.g. public health
- Potentially a mix of academic and non- academic doctors

The initial panels recruited fulfilled the characteristics we aimed to achieve as follows:

Public/Patients

Age range:	early 20s – late 60s
Gender:	8 males; 12 females
Ethnicity:	18 white; 2 minority
Employment:	Various fields including retired academic, actor, artist, charity worker, clerical workers, farm worker, human science student, research assistants, self-employed, trainee solicitor, sound engineer, teacher, tree surgeon, writer and unemployed
Residence:	Mixed rural and urban

Nurses:

Age range:	20s - 50s
Gender:	4 males; 11 females
Ethnicity:	14 white; 1 minority
Experience:	8 – 43 years' experience in various fields including general nursing care, clinical practice educator, community care, nurse specialists in child psychiatry and adolescent and child development and clinical team leaders
Residence:	Mixed rural and urban

Doctors:

Age range:	20s – 50s
Gender:	11 males; 4 females
Ethnicity:	11 white; 4 minority
Experience:	1 – 30+ years post-qualification ranging from foundation year 1 – senior consultant. Specialist fields include accident and emergency medicine, paediatrics, child and adolescent psychiatry, old age psychiatry and general practice
Residence:	Mixed rural and urban

Allied Health Professionals:

Age range:	20 - 30s
Gender:	5 females
Ethnicity:	5 white
Experience:	7 – 15 years specialist experience as physiotherapists, dieticians, occupational therapists
Residence:	Urban

Responsible Officers/Medical Directors:

Age range:	40s – 60s
Gender:	7 males
Ethnicity:	6 white; 1 minority
Experience:	24 – 30+ years as ROs and/or MDs, CEOs. Specialist fields include, orthopaedics and trauma, surgery, pain management, anaesthetics and general practice
Residence:	Mixed rural and urban

Overall panel characteristics:

Age range:	early 20s – late 60s
Gender:	30 males; 32 females
Ethnicity:	54 white; 8 minority
Experience:	1 – 40+ years
Residence:	Mixed rural and urban

3.2.3. Panel sites

Based on expressed willingness to participate in the research on the part of the participants, the following regions/cities were identified as panel sites:

Doctors:
Bury St. Edmunds (to include doctors working in East Anglia and the East Midlands)
Dumfries (to include doctors working in South-West Scotland)
London (to include doctors working in London and Southern England)

Nurses:
Belfast (to include nurses working in rural and urban areas of Northern Ireland)
Bury St. Edmunds (to include nurses working in East Anglia and the East Midlands)
London (to include nurses working in London and Southern England)

Public/Patients:
Caernarfon (to include members of the public living in rural and urban areas of North Wales)
London (to include members of the public living in London and the South-East)
Newcastle (to include members of the public living in rural and urban areas of the North-East)

Allied Health Professionals:
London (to include AHPs working in London and Southern England)

Responsible Officers/Medical Directors:
Birmingham (to include ROs and MDs working in the East and West Midlands, Yorkshire and Lancashire)

3.2.4. Location and dates of initial stakeholder panels

A series of initial stakeholder panels was convened as shown in Table 1.

Date (2012)	Location	Participants	Number
16/17 May	Caernarfon, Wales	Patients	7
28/29 May	London	Nurses	5
30/31 May	London	Patients	6
8/9 June	Newcastle	Patients	7
15/16 June	Dumfries, Scotland	Doctors	5
21/22 June	Bury St. Edmunds	Doctors	5
28/29 June	Bury St. Edmunds	Nurses	5
19/20 June	London	Allied Health Professionals	5
31July/1August	Belfast, Northern Ireland	Nurses	5
20/21 August	London	Doctors	5
13 September	Birmingham	ROs/Medical Directors	7

Table 1: Initial stakeholder panels

3.2.5. Procedure

A framework for conducting the panel discussions, including guiding questions, was developed. The first panel for public/patients was held on 16 and 17 May in Caernarfon, North Wales. Seven participants, ranging in age from early twenties to late sixties, and with diverse backgrounds were recruited. Five were fluent in English and Welsh with four of the panel stating that Welsh was the language spoken at home. As this was the first time the materials had been used, the researchers saw this as an opportunity not only to obtain data regarding optimum IELTS levels but also as a chance to refine the guiding framework for maximum practical efficiency in the remaining sessions. The sessions were conducted as follows:

1. The lead moderator introduced the project – background, aims and objectives. Questions from panel participants relating to the project were responded to as appropriate.
2. Research consent, non-disclosure and bank account forms were distributed and signed forms collected (participants had been informed in advance of the information required for the forms and requested to bring it with them).
3. The IMG/EEA distinction was explained to participants and the following scenario (originally provided by the GMC) was outlined to further explain the intricacies and subtleties of the IMG/EEA distinction:

   > "EEA doctors would include EEA nationals, Swiss nationals and those entitled to be treated as such (by virtue of an EC right). An EC right can be demonstrated in a number of ways. One example would be a Chinese doctor married to a French national who is working in Belgium. This doctor would then be treated in the same way as an EEA national."

   The panel then discussed the IMG/EEA distinction issue to decide if both should be treated as a single population with the same language requirements or as separate

populations with different language requirements (Research Question 2). Discussions, both of this issue and of all further panel deliberations, were recorded.

4. Immediately prior to considering each skill individually, participants were asked to brainstorm the language characteristics that they believe a minimally competent candidate (MCC) should possess for each skill. These characteristics were summarized as 'can do' statements and supplemented by IELTS descriptors and CEFR 'can do' statements (see Appendix 1 for 'can do' statements developed for each skill).
5. Participants were then asked to judge a series of oral performances/ writing samples/ reading and listening task responses and make individual initial decisions as to acceptability. The four skills were considered separately, as follows:
    - Two reading tests were completed (3 reading texts and 40 questions for each test) and each participant stated how many questions from each test a MCC should answer correctly. The totals were averaged for each test and converted to a band score according to the raw score to band conversion tables supplied by Cambridge ESOL. A composite Band score was then produced based on responses to both tests.
    - Twelve writing scripts were studied and ranked from best to worst. Each participant stated which script represented the minimum acceptable level of writing ability. Participants were then told what band each script represented and asked to discuss the minimum band they would accept from a MCC. Each script was then ranked on a 5-band scale of x (not acceptable) - ✓✓ (better than acceptable). For details of the scale, see i) below.
    - Two listening tests were completed and each participant stated how many questions from each test a MCC should answer correctly. The totals were averaged and converted to a band score according to the raw score to band conversion tables supplied by Cambridge ESOL. A composite Band score was then produced based on responses to both tests.
    - Sixteen speech samples were listened to, discussed and rated as acceptable or not on a scale of x (not acceptable) - ✓✓ (better than acceptable). Participants were then told what band each speech sample represented and asked to discuss the minimum band they would accept from a MCC.
6. After each skill had been discussed and the minimum level confirmed, participants were asked to consider how easy they had found it to match the task requirements of the test to the 'can do' statements they had developed for that skill (Research Question 3).
7. Following completion of all papers and discussion of the tasks in relation to the 'can do' statements, panel participants were asked to confirm their original decisions for each skill and, finally, to determine an overall minimally acceptable band level (Research Question 1).

8. A discussion then ensued between the researchers and the participants regarding the work-load and concentration levels that had been required of them to complete the two days.

Following these discussions, the procedures were modified to reduce the required workload.

- In order to better accommodate the very different reading speeds of participants, the number of writing scripts was reduced from twelve to eight, but continued to include one writing exemplar from each band level.
- The number of speech samples was reduced from sixteen to twelve and, in order to reduce the time spent by judges when listening to performances at different IELTS levels, audio recordings of interviews were edited in order to provide a minimum of 5 – 7 minute extracts of candidate speaking time. All edited speech samples included initial information regarding the candidates' name and country of origin, the 1 – 2 minute monologue on a range of different topics and 2 – 3 further follow-up exchanges based on the long turn topic.
- As with the writing test, in order to better take into account the different reading speeds of the panel participants, one reading text and associated questions were studied as a practice test and only one complete reading test was formally judged.

All subsequent initial stakeholder panels then followed the amended format. Points 1 - 4 and Points 6 - 7 were completed as outlined above. However judgments of individual skills were made as follows:

Receptive skills – reading and listening

a) Participants were given one text from one of the **reading** tests and the questions associated with that text. They were asked to read the text and questions in order to familiarise themselves with the format, task types and general design of the test, something which the IELTS website tells candidates is essential for successful performance. It was explained to participants that candidates taking the IELTS test would probably have practised it, may have taken training courses and would generally be very familiar with the requirements. After the panel had studied the reading text and questions, the lead moderator introduced the full range of task types that might be encountered in an IELTS reading test.

b) Participants then completed the second reading test (3 texts and 40 questions) and determined how many questions a minimally competent doctor should be able to answer correctly.

c) Participants were given a reading response sheet and asked to mark each question separately then total their responses at the end, giving a score out of forty for reading.

d) The moderators then added each participant's scores, averaged them and converted the average score obtained to a band according to the raw scores to bands conversion table provided by Cambridge ESOL.

e) The panel members were informed of the average band they had decided on for the reading skill and invited to discuss this and confirm their decision.

f) The procedure was repeated for the **listening** test with the important exception of the first point (point a). Since the listening test is done in 'real time' and is therefore not dependent on individual reading speed, the panels were given the first listening test to complete as a 'practice test' in order to familiarise themselves with the format, task types and general design of the test (in line with general advice to candidates). After the panel had completed the first listening test (4 sections, 40 questions), the lead moderator introduced the full range of task types that might be encountered in an IELTS listening test. The second listening test was then completed and the procedures in points b – e (relating to the reading test) were then followed.

Productive skills –writing and speaking

g) For the **writing** test, participants were asked to rank order 8 scripts which had been standardized by Cambridge ESOL and rated ranging from Band 5 – Band 8.5

h) Participants were first asked to determine whether the scripts were acceptable or not as writing from a doctor

i) They were then asked to look at the scripts a further time and complete a writing response sheet giving them a rating as follows:

5 = very good writing, better than acceptable (✓✓)
4 = acceptable writing for a doctor (✓)
3 = borderline acceptable (✓?)
2 = borderline not acceptable (x?)
1 = not acceptable writing for a doctor (x)

j) For the **speaking** test, participants listened to individual speaking tests from 12 candidates who had been given a standardized rating by Cambridge ESOL ranging from Band 5 – Band 8.5.

k) Participants were then asked to determine if each candidates' speech was acceptable or not for a doctor.

l) They were then asked to rate each candidates as follows:

5 = very good, better than acceptable (✓✓)
4 = acceptable speech for a doctor (✓)
3 = borderline acceptable (✓?)
2 = borderline not acceptable (x?)
1 = not acceptable speech for a doctor (x)

m) Initial individual decisions for the writing and speaking tests were analysed by averaging the score given to each writing and speech sample. Those in the 4 or over range were definitely acceptable. Those in the 3 - 4 range were almost, but not definitely acceptable so the panel members were asked to discuss their decisions and come to a collective agreement as to the acceptability of each writing and speech sample and the definition of the

writing/speaking competence of a minimally competent doctor. Speech samples for some candidates were replayed, as requested.

n) These were then reorganised as Band levels according to the criteria supplied by Cambridge ESOL and the panel (with reference to IELTS writing/speaking descriptors as necessary) was informed of the Band they had agreed on and asked to discuss it and confirm their decision.

Panels were also asked to consider the match between the task demands for each skills and the respective 'can do' statements; collective agreement was then sought from each panel as to the minimum IELTS level requirement for each paper and overall as in Points 6 and 7 of the original outline for the Welsh patients panel (above).

3.3. Final panel

Following completion of the initial stakeholders' panels, all judgments and data were analysed and comments on each of the skills discretely and overall were collated. Representatives from each category of panel were invited to participate in a final confirmatory panel in London on 12 October. Participants in the final panel were offered accommodation and travelling expenses, as necessary, and were each paid an incentive fee of £310. Two patients, two doctors, two nurses, one allied health professional and one RO/MD participated in the final panel, which was organised as follows:

1. The convener reminded the panel of the basic aims and objectives of the research and reviewed the methodology of the study.
2. Participants were presented with summaries of 'can do' statements for all skills (see Appendix 1) and summarized comments from initial panels on each of the IELTS skills tests (see Appendix 2).
3. Summaries of qualitative judgments for each skill were presented by panel (individual panels) and by category of panels (patients', doctors', nurses' judgments averaged) and discussed; this discussion, and all subsequent discussions, was recorded for later transcription and review.
4. Quantitative analyses were presented (descriptive statistics including averaged and ranges of scores for Reading and Listening, converted to Band scores, plus means of scores ranging from 1 – 5 which had been allocated to speech samples and writing exemplars) and discussed.
5. Decisions were made regarding appropriate levels to recommend for each skill and overall.
6. The appropriacy of the IELTS as a screening test for overseas doctors prior to registration for PLAB Part 1 was discussed and decisions on recommendations were made.
7. The IMG-EEA distinction was discussed and decisions on recommendations were made.
8. Evidence relating to the testing systems in place for other UK professionals, other UK regulators and overseas medical regulators was presented and discussed (briefly).

9. Overall recommendations to be offered to the GMC were noted, related back to the final panel, further discussed and confirmed.
10. Once the first draft of the report had been written, the executive summary and recommendations therein were circulated to the final panel for confirmation that the recommendations reflected the panel's decisions.

4. Results and discussion of findings from initial stakeholders' panels

In this section the three questions arising from the literature review will be addressed in a slightly different order from that presented in Section 2. First, we will consider results from the deliberations of the initial panels as to the appropriate IELTS levels for each skill and overall (Question 2). At the same time, we will look at comments about the IELTS test made by the initial panels to determine if the IELTS is an appropriate test for screening of overseas doctors (Question 1). Finally, we will consider comments from the initial panels with relation to the situation between IMG applicants to the GMC register and those from the EEA or holding EC rights (question 3).

4.1. Subjective judgments from the initial stakeholder panels

In order to bring the judgments from the Caernafon patients panel into line with the other panels, data obtained were amended to include only the same writing papers and speech samples that the subsequent ten panels assessed; also, for the same reason, only the results from the second reading and listening tests were included in the analysis.

Recommended IELTS bands for each skill and overall based on the subjective decisions of each individual panel are summarised in Table 2.

Panel	Reading	Writing	Listening	Speaking	Overall
Caernafon patients	Band 7.5	Band 7	Band 8	Band 7.5	Band 7.5
London patients	Band 8.5	Band 8	Band 8.5	Band 8	Band 8
Newcastle patients	Band 8	Band 7	Band 8.5	Band 8	Band 8
Belfast nurses	Band 8	Band 7	Band 8.5	Band 7.5	Band 7.5
East Anglia nurses	Band 8	Band 8	Band 8.5	Band 8.5	Band 8
London nurses	Band 8.5	Band 8	Band 9	Band 8	Band 8
Dumfries doctors	Band 8	Band 9	Band 9	Band 7	Band 8
East Anglia doctors	Band 7	Band 6.5	Band 7.5	Band 7	Band 7
London doctors	Band 7	Band 7	Band 8.5	Band 7.5	Band 7.5
A.H. Professionals	Band 7	Band 7.5	Band 8.5	Band 7.5	Band 7.5
ROs/Medical directors	Band 7.5	Band 7.5	Band 8.5	Band 7.5	Band 7.5

Overall by panels	**Band 7.5** *(7.73)* range 7–8.5	**Band 7.5** range 6.5-9	**Band 8.5** *(8.45)* range 7.5-9	**Band 7.5** *(7.63)* range 7-8.5	**Band 7.5** *(7.68)* range 7-8

Table 2: Subjective collective decisions of individual panels – bands for each skill and overall (actual mean in brackets)

As can be seen, there is much variation between the individual panels with only the judgment of overall band score required being fairly consistent across groups. London patients, East Anglian and London nurses and Dumfries doctors (with the exception of speaking) make the harshest judgments and East Anglian doctors are, in general, fairly lenient.

However, when judgments of recommended band scores are averaged by category of panel, i.e. for public/patients, doctors, and nurses as groups, (allied health professionals and responsible officers remain the same as there was only one panel for each of them) the results are much more consistent across categories, those for nurses and patients being identical. Results by category of panels are shown in Table 3.

Categories	**Reading**	**Writing**	**Listening**	**Speaking**	**Overall**
Patients	Band 8 *(8.16)*	Band 7.5 *(7.33)*	Band 8.5 *(8.66)*	Band 8 *(7.83)*	Band 8 *(7.83)*
Nurses	Band 8 *(8.16)*	Band 7.5 *(7.66)*	Band 8.5 *(8.66)*	Band 8	Band 8 *(7.83)*
Doctors	Band 7.5 *(7.33)*	Band 7.5	Band 8.5 *(8.33)*	Band 7 *(7.16)*	Band 7.5
AHPs	Band 7	Band 7.5	Band 8.5	Band 7.5	Band 7.5
ROs/MDs	Band 7.5	Band 7.5	Band 8.5	Band 7.5	Band 7.5
Overall by category	**Band 7.5** *(7.63)* range 7 - 8	**Band 7.5** *(7.49)*	**Band 8.5** *(8.53)*	**Band 7.5** *(7.59)* range 7 - 8	**Band 7.5** *(7.63)* range 7.5 - 8

Table 3: Subjective collective decisions of categories of panels – bands for each skill and overall (actual means in brackets)

As can be seen from Table 3, every category of panel has opted for a band score of 8.5 for **listening**. Although this is an exceptionally high requirement, it is reflective of a number of comments made about the listening test including[7]:

AHP It's a good test of general listening ability but it's too clean, it's not listening in the real world and some of it is too artificial and structured.

AHP Doctors need to be able to pick up information from much more messy discourse with background noise.

[7] Comments are attributed to participants as follows: D = doctors; N = nurses; P = patients; AHP = Allied Health Professionals; RO = ROs/MDs

RO *If they can't do this test, they won't be able to manage in A&E where they've got anxious relatives yelling at them and patients who are pretty incoherent.*

D *It's a listening test but it's not at all appropriate as a measure for assessing whether a doctor can cope in a clinical setting of any description.*

D *The audio is a lot easier to understand than a lot of hospitals, in a busy A & E – and regional accents and colloquialism, none of that. And people don't speak in such neat sentences.*

D *The listening test is just farcical.*

N *The questions run chronologically which makes it very, very easy.*

N *They speak very clearly, the pronunciation is very, very clear and it runs chronologically so if we are going to assess someone's ability to listen to people in real life, I don't think it's a particularly useful tool to assess people against what we are actually expecting them to do.*

P *In some respects this is quite a good test of some of the things we think doctors should do but because of the tempo and the guidance that is given, if a doctor is going to be able to function in a genuine medical setting they would need to get all of the questions right.*

P *This is too easy, basically. The listening seemed a level below the reading.*

Comments are organised by category of participant and close examination of the comments relating to the listening test show that there is very little difference in the types of comments made about each of the skills by each of the categories. Although all categories commented that it appears to be a good test of general listening ability, they also all questioned its appropriacy as a means of providing evidence of a doctor's listening ability, highlighting features of the listening test such as the clarity and lack of speed of articulation, lack of variety of accents and colloquialisms and especially the cleanness of presentation and lack of background noise. There were also comments from every panel as to the lack of realistic and difficult telephone conversations, which are considered an essential listening skill for a doctor.

P *There should be other tasks like an emotional situation, someone describing something in tears, on the telephone, with background noise.*

P *The telephone conversation was very clear compared to normal.*

D *I think in practical things that could be done, they could have a telephone consultation, just things like that, that would be a way of having a test that examines all the various areas of communication skills that we have been looking at today in an integrated fashion and in a context that is relevant to the job that they are going to be doing in this country,*

RO *They need to listen over the phone. Listening face-to-face and listening over the telephone are two important and sometimes different methods of communication.*

RO *The telephone is difficult, it's harder than face-to-face, isn't it. Telephone conversation with an agitated relative.*

RO *They need a test like the call handlers 999 ambulance services. Don't they have a test with agitated, stressed people? They have to be able, as part of their test, to get accurately location, symptoms, with an awful lot of background noise. One scenario they have got is a drunk person who is on the phone and they are tested on the accuracy of what they can get from the conversation.*

N *So carry out a phone conversation, cutting out interruptions, and at the same time processing the thoughts into something more real. So you're not just listening and then get five minutes to recap on the information and necessary questions that you might have for the person on the other end of the phone. You are listening and the whole time you are making decisions on whether or not you need to bring the patient over and what test needs to be – things need to be pretty much summarised by the time you put the phone down.*

(For a full range of all the comments made about the different skills in the IELTS test, see Appendix 2.)

As can also be seen from Table 2, every category of panel also opted for an identical band score of 7.5 for the **writing** test. However, this might be an artefact of averaging the categories as the range of judgments for writing was from band 6.5 to band 9. This is reflected clearly in the almost diametrically opposed comments from different categories such as:

AHP *It's quite a good indicator of vocabulary and structure and grammar and logic so I'm happy with the task.*

N *What I liked about this task is that they were having to consider various arguments and put them together and then come up with an opinion.*

N *I think it weeds out some of the poorer candidates so it's quite good from that point of view.*

N *It's an adequate task to get information.*

P *These aren't the kinds of writing we expect doctors to be able to do.*

P *We need different, more specific information.*

P *It doesn't actually reflect the sorts of writing that we think doctors have to do.*

D *It doesn't give me enough information whether a doctor could write what they need to write.*

D *This test doesn't enable is to predict anything to do with doctors' writing skills, the sorts of skills a doctor is really going to need.*

D *This is not an appropriate test for doctors and shouldn't be used.*

RO *I don't feel strongly about it because for me they are arbitrary numbers.*

It seems from the comments that AHPs and nurses are reasonably happy that the writing tests provides appropriate evidence of a doctor's ability to write whereas patients and doctors were distinctly less impressed, with one panel of doctors insisting that the test should not be used. ROs had no comments to make about the writing test itself but felt

that the descriptors of the different bands in the writing band scale were random and subjective.

Decisions for an appropriate band score for **reading** were quite consistent with a range of 7 – 8, with only one category (in fact only a single panel) opting for 7. It is likely that decisions were taken to allow scope for mistakes on the reading test as it is was clearly perceived to be rather difficult. Interestingly, the comments about the reading test were quite diverse with some thinking it a very good test and others thinking it not really appropriate for doctors:

N *It's a good general test but there should be some tasks that are more geared to the skills a doctor needs.*

N *I don't think it gives you enough information as to whether a doctor has sufficient reading skills to practise in this country because it's missing critical skills areas they need.*

N *This is a very good test to tell if someone can read but I think there's scope to make it more specific, to fine tune it.*

P *It's a good test of reading ability in general but it's not sufficient to tell us whether a doctor can read the sorts of things they need to read.*

D *I think they are very fair questions. Sometimes the answers are a little bit more subtle than they would be in a guideline or hospital policy.*

D *I think it's a reasonably good way to assess reading in English.*

D *I don't think it's testing what it needs to test.*

AHP *It's a good test of your interpretation of implied things and I think it tests your ability to weed things out of a text.*

AHP *It doesn't tell us whether a doctor is going to be able to read all the things we determined they should be able to read.*

However, unlike with the listening test, none of the panels judged a band score higher than 8 was required for reasons probably reflected in the following comments:

RO *It's not easy.*

RO *I think we all thought it was quite challenging and that is reassuring.*

N *It's quite difficult.*

The final comment originally relating to listening is worth reiterating here also:

P *The listening seemed a level below the reading.*

As Eckes (2009:10) notes, assessing speaking performance is extremely complex, particularly in direct speaking tests, due to the interaction of various facets identified in studies by Berry (2007), Bonk and Ockey (2003), Brown (2005), O'Sullivan (2008) and van Moere (2006), *inter alia*. In the current study, the speech samples proved quite difficult to judge, partly because they were all re-recorded on audio only (for logistical reasons), whereas on the actual **speaking** test, there would have been face-to-face interaction making it easier to judge whether, for example, hesitation was due to

genuinely thinking about the topic or searching for vocabulary, as reflected in this comment:

N *It's difficult to gauge whether someone is hesitating because they can't think of the right words or because they can't think of anything to say.*

The range is quite narrow, from Band 7 to Band 8. Most panels initially included a Band 7 candidate because she has an almost native-speaker-like American accent. However on closer listening, she was rejected by all but two of the panels for the following reasons:

N *She answered the questions appropriately but her vocabulary was reasonably limited in the regard that she talks like a TV American teenager. She keeps saying 'like' and 'you know' which I think would be quite distracting for people who were trying to listen.*

N *She was repetitive and also she was giggling nervously throughout rather than because of humour.*

D *She superficially sounds alright and then you realize that she can't really explain anything in great detail or depth.*

P *I don't think she could cope in a hospital situation.*

Conversely, the Band 7.5 candidate was originally rejected by most panels on the grounds that her accent is almost incomprehensible as she has a very strong, guttural Polish accent. However, several panellists commented that her grammar and vocabulary were good and, after a second listening, she was acceptable to all but the patients and nurses. Interestingly, several panellists commented that she was more comprehensible than some of the actual overseas doctors they currently work with, which may have made them treat her more leniently than would otherwise have been the case.

Regarding the actual speaking test, as with the other skills, panellists' reactions were mixed regarding its appropriacy for doctors but, in general, were fairly positive:

N *I thought it was quite a good test, really. The good ones really stood out. I thought them having to present a topic was a good test of organizational ability in ordering your thoughts.*

P *From the sorts of questions and the sorts of tasks, we couldn't say whether someone could be a doctor.*

AHP *It's quite a good test because there is enough opportunity to display your language ability.*

N *Getting your point across when there is more than one person talking across a cacophony of noise may be more challenging.*

AHP *I think it's good. I think it gives you a good idea of their fluidity of speaking, how they structure their responses and their sentences and their grammar and their vocabulary.*

With the exception of ROs and nurses, who thought it was appropriate to test each skill discretely, the other three categories of panels commented that it would be more realistic to test integrated skills:

AHP I was thinking they could do reading and speaking, the speaking task should involve explaining something that was written.

AHP If a doctor gets results of, gets a written report from an x-ray or something, they have to read and then go on and explain in non-technical terms to a patient, that's quite a high skill. I think they need to read that information, extract it and then be able to explain it in simple terms to a patient.

AHP In the test as it stands you are prompted to look for certain answers but if you are given a script and you have to extract it yourself, that is a whole other skill.

AHP I think the listening was good because it involved a bit of writing so you had to listen and then write down the points.

AHP If they had a piece of writing and they listened to something and then they had to write it down, I think it should all be holistic.

P I think they should all be together.

P There is a case to have some aspects of writing to be assessed separately and some aspects of reading and then all together.

P There are crossovers everywhere because you can say listening and reading, listening to the patient then you have to go and read and learn stuff and then you have to write about that and having a whole integrated thing would be better.

P Nine times out of ten with GPs you are listening and listening and writing down then you come up with what you know.

P There should be a mini cross-over test for all.

D It makes sense to integrate a lot of the different skills together.

D They should do something with what they've heard that is completely different like listening to something then writing a referral.

D I think the writing and the speaking are being artificially divided.

D I think you can have multiple different outputs from one input and one example might be from a ward round input, one might be to write a clinic letter, one might be to write a communication as a GP to a patient.

D I think part of the trouble is we are trying to assess each part of the language separately whereas they are all interlinked and you can't assess the full complexity of one aspect of it without assessing the other components of the language at the same time.

One final comment, which was originally made about the speaking test, but could in fact have been made in relation to any skill, or to the test overall, is worth offering here:

D Part of the problem is we are using a test that was originally designed for one purpose, for a completely different purpose, so it's never going to be perfect.

4.2. Quantitative analysis of scores allocated for Reading and Listening

As mentioned in Section 2, both the IELTS receptive skills tests (reading and listening) consist of a total of 40 questions each. Panels were asked to specify how many of those questions a candidate should be able to answer correctly and give a score out of 40. Descriptive statistics for reading and listening were calculated using IBM SPSS Version 19.

4.2.1. Reading

The Reading test has a total of 40 questions. The overall mean for all panels for the reading test is shown in Table 4. According to the score to band conversion table provided by Cambridge ESOL for this specific test, a mean of 35 (out of a total of 40) represents a Band Score of 8, which is half a band higher than the overall subjective judgment for Reading of 7.5 shown in Table 2.

Reading

Mean	N	Std. Deviation
35.2258	62	3.93979

Table 4. Overall mean for Reading.

Means for reading for each **panel** separately are shown in Table 5.

Reading

Panel	Mean	N	Std. Deviation
LP	35.1667	6	1.32916
NP	35.1429	7	4.29839
CP	33.2857	7	3.25137
EAN	36.6000	5	2.70185
BN	40.0000	5	.00000
LN	39.4000	5	.89443
EAD	31.4000	5	2.88097
LD	33.0000	5	5.29150
DD	34.4000	5	1.34164
AHP	30.6000	5	3.20936
RO-MD	38.1429	7	2.34013
Total	35.2258	62	3.93979

Table 5. Mean scores for each panel for Reading[8]

[8] Key: LP = London patients; NP = Newcastle patients; CP = Caernafon patients; EAN = East Anglia nurses; BN = Belfast nurses; LN = London nurses; EAD = East Anglia doctors; LD = London doctors; DD

As can be seen, there is quite a considerable difference (almost 25%) in scores each panel thought should be able to be answered correctly ranging from 30.6 to 40 (Band 7 – Band 9). A comparison of the original subjective judgments and subsequent band score equivalents awarded by each panel is shown in Table 6. Of the 11 panels, 5 thought more questions should be answered correctly, resulting in a higher Band level than they gave in their subjective decisions; 5 panels gave identical ratings both as scores and subjective judgments and only 1 thought fewer questions should be answered correctly than their subjective judgment suggested.

Panel	**Subjective judgment Band**	**Score equivalent Band**	**Difference**
LP	Band 8.5	Band 7.5	+ .5
NP	Band 8	Band 8.5	=
CP	Band 7.5	Band 7.5	=
EAN	Band 8	Band 8	=
BN	Band 8	Band 9	+ .5
LN	Band 8.5	Band 9	+ .5
EAD	Band 7	Band 7	=
LD	Band 7	Band 7.5	+ .5
DD	Band 8	Band 7.5	- .5
AHP	Band 7	Band 7	=
RO-MD	Band 7.5	Band 8	+ .5

Table 6: Comparison of original subjective judgments and subsequent score equivalent bands for Reading by panel.

Appendix 4 presents a complete Analysis of Variance (ANOVA) with *post hoc* comparisons of both reading and listening scores awarded by panel, category, gender, ethnicity and age.

Analysis of variance of scores awarded for reading by **panels** show that significant differences exist between the scores awarded between the following panels: AHP and BN, AHP and LN, EAD and BN. No other differences between the panels were significant.

Means scores for reading for each **category** are shown in Table 7. Table 8 shows the differences between each category's original subjective judgment and subsequent score equivalent band awarded. Doctors and AHPs gave identical ratings both as subjective judgments and score equivalents, nurses and ROs thought more questions should be answered correctly, resulting in a higher Band level than they gave in their subjective decisions; and only 1 thought fewer questions should be answered correctly than their subjective judgments suggested.

= Dumfries doctors; AHP = Allied Health Professionals; RO-MD = Responsible Officers – Medical Directors

.
Analysis of variance of scores awarded for reading by categories show that significant differences exist between the scores awarded between the following categories: nurses and patients, nurses and doctors, nurses and AHPs, RO-MDs and doctors, RO-MDs and AHPs. No other differences between the categories were significant.

Reading

Category	Mean	N	Std. Deviation
patients	34.5000	20	3.23631
nurses	38.6667	15	2.16025
doctors	32.9333	15	3.53486
allied health professionals	30.6000	5	3.20936
RO/employers	38.1429	7	2.34013
Total	35.2258	62	3.93979

Table 7: Mean scores for each category for Reading.

Category	**Subjective judgment Band**	**Score equivalent Band**	**Difference**
Patients	Band 8	Band 7.5	- 1
Nurses	Band 8	Band 8.5	+ .5
Doctors	Band 7.5	Band 7.5	=
AHPs	Band 7	Band 7	=
RO/MDs	Band 7.5	Band 8.5	+ 1

Table 8: Comparison of original subjective judgments and subsequent score equivalent bands for Reading.

ANOVA of differences of scores awarded for reading based on **gender**, **ethnicity** and **age** revealed no significant differences between groups.

4.2.2. Listening

The listening test also consists of 40 questions. The overall mean for all panels for the listening test is shown in Table 9. According to the score to band conversion table provided by Cambridge ESOL, a mean of 37 (out of a total possible of 40) on this test represents a Band Score of 8.5, which is identical to the overall subjective judgment for Listening shown in Table 2.

Listening

Mean	N	Std. Deviation
37.0000	62	2.50900

Table 9: overall mean for Listening.

Means for Listening for each panel separately are shown in Table 10.

Listening

Panel	Mean	N	Std. Deviation
LP	36.6667	6	2.73252
NP	35.8571	7	1.21499
CP	36.4286	7	1.39728
EAN	37.2000	5	1.92354
BN	37.2000	5	2.68328
LN	39.4000	5	.54772
EAD	32.0000	5	3.39116
LD	37.0000	5	.00000
DD	39.2000	5	1.09545
AHP	37.0000	5	1.22474
RO-MD	39.0000	7	1.00000
Total	37.0000	62	2.50900

Table 10: Mean scores for each panel for Listening.

As can be seen, there is quite a considerable difference (almost 20%) in scores each panel thought should be able to be answered correctly ranging from 32 to 39 (equivalent to Band 7.5 – Band 9). A comparison of the original subjective judgments and subsequent band score equivalents awarded by each panel for listening is shown in Table 11. The scores are remarkably consistent with the judgments, with 9 of the 11 panels giving identical ratings both as scores and subjective judgments; Caernafon patients and RO-MDs thought more questions should be answered correctly than their subjective judgments warranted.

Panel	**Subjective judgment Band**	**Score equivalent Band**	**Difference**
LP	Band 8.5	Band 8.5	=
NP	Band 8.5	Band 8.5	=
CP	Band 8	Band 8.5	+ .5
EAN	Band 8.5	Band 8.5	=
BN	Band 8.5	Band 8.5	=
LN	Band 9	Band 9	=
EAD	Band 7.5	Band 7.5	=
LD	Band 8.5	Band 8.5	=
DD	Band 9	Band 9	=
AHP	Band 8.5	Band 8.5	=
RO-MD	Band 8.5	Band 9	+ .5

Table 11: Comparison of original subjective judgments and subsequent score equivalent bands for Listening by panel.

Analysis of variance with *post hoc* comparisons of scores awarded for listening by panels show that significant differences exist between the scores awarded between East Anglia

doctors (EAD) and EAN, BN, LN, DD and RO-MDs. No other differences between the panels were significant.

Means for listening for each **category** are shown in Table 12. As can be seen, when comparisons are made with their initial judgments (Table 13), patients', nurses' and allied health professionals' judgments were identical; doctors underestimated how many questions should be answered correctly and ROs overestimated, both by the equivalent of .5 of a band.

Listening

Category	Mean	N	Std. Deviation
patients	36.3000	20	1.78001
nurses	37.9333	15	2.08624
doctors	36.0667	15	3.65409
allied health professionals	37.0000	5	1.22474
RO/employers	39.0000	7	1.00000
Total	37.0000	62	2.50900

Table 12: Mean scores for each category for Listening

Category	**Subjective judgment Band**	**Score equivalent Band**	**Difference**
Patients	Band 8.5	Band 8.5	=
Nurses	Band 8.5	Band 8.5	=
Doctors	Band 8.5	Band 8	- .5
AHPs	Band 8.5	Band 8.5	=
RO/MDs	Band 8.5	Band 9	+ .5

Table 13: Comparison of original subjective judgments and subsequent score equivalent bands for Listening.

Analysis of variance with *post hoc* comparisons shows that no significant differences exist between scores awarded for listening by categories. ANOVA of differences of scores awarded for listening based on **gender**, **ethnicity** and **age** also revealed no significant differences between groups.

4.3. Quantitative analysis of scores allocated for Writing and Speaking

In order to include the decisions of the Caernafon patients in the quantitative analysis, ratings of x - ✓✓ were converted to a numerical scale, 1 – 5, as discussed in Section 3.2.5. i) and 3.2.5.l), above.

4.3.1. Writing

As described in Section 3.2.5.i, above, each script was ranked and then allocated a score as follows:

5 = very good writing, better than acceptable (✓✓)
4 = acceptable writing for a doctor (✓)
3 = borderline acceptable (✓?)
2 = borderline not acceptable (x?)
1 = not acceptable writing for a doctor (x)

It is important to emphasise that any numbers allocated to respective subjective decisions are, by definition, arbitrary. What is crucial to an understanding of the quantitative analysis is that having allocated the scores on a scale of 1 – 5, each participant gave their score based on how the arbitrary number reflected the description accompanying it. In other words, if a participant was confident that the piece of writing was clearly acceptable for a doctor, that piece of writing was given 4; however if a problem of whatever nature was perceived and the piece of writing was considered to be almost but not quite acceptable, then it was given 3. It will be remembered from Section 3.2.5. that after each piece of writing had been ranked and awarded a score borderline decisions were reviewed and reconsidered so a final decision to award a score of 3 rather than 4 was only taken after considerable reflection. As any piece of writing which, when scores were averaged, remained in the 3 - 4 range had to be considered borderline, a minimum average of 4 was therefore required to ensure a clear decision of acceptability.

Table 14 shows the overall mean scores for each script. As can be seen from Table 14, only Script 7 achieves an overall mean of 4 or over, which is the minimal accepted score; Script 7 represents Band 8.5.

	Script1	Script2	Script3	Script4	Script5	Script6	Script7	Script8
Mean	1.63	1.90	3.21	2.18	3.10	3.68	4.68	3.82
N	62	62	62	62	62	62	62	62
Std. Deviation	.910	.987	1.243	1.109	1.020	.937	.505	.713

Table 14: Overall mean scores for each writing exemplar.

Table 15 shows the mean scores allocated to each script by each panel. As can be seen from Table 15, Script 7 was acceptable to all panels. Script 8, representing a band-score level of Band 7 was acceptable to 5 panels, but not overall; Script 6, representing a band-score level of Band 6 was acceptable to 3 panels, but not overall; somewhat surprisingly, Script 5, representing a band-score level of Band 8 was only acceptable to 1 panel. These findings highlight the difficulties encountered by the participants in judging the writing samples; these findings are also consistent with the contradictory findings for writing which emerged from the 2004 study (Banerjee, 2004:A7).

Table 16 shows mean scores for each writing exemplar for each category. As can be seen from Table 14, Script 7, representing a band-score level of Band 8.5, is acceptable for all categories; Script 8, representing a band-score level of Band 7, is acceptable to nurses and ROs, but not overall. When comparisons are made with their initial judgments (Table 17), no category produces the same band judgment and band score equivalent. Patients,

doctors and AHPs have higher band score equivalents than their subjective judgments whereas the reverse is true for nurses and ROs/MDs.

Panel		Script1	Script2	Script3	Script4	Script5	Script6	Script7	Script8
LP	Mean	1.50	2.17	3.00	2.50	2.50	2.83	5.00	3.67
	N	6	6	6	6	6	6	6	6
	Std. Deviation	.837	.983	1.095	1.049	1.643	1.169	.000	.816
NP	Mean	1.14	2.00	2.71	2.71	3.00	4.14	4.71	3.71
	N	7	7	7	7	7	7	7	7
	Std. Deviation	.378	1.000	1.254	.951	1.000	.690	.488	.488
CP	Mean	1.29	1.71	2.71	1.86	2.57	3.14	4.57	3.86
	N	7	7	7	7	7	7	7	7
	Std. Deviation	.488	.756	1.113	1.069	.787	.900	.535	.900
EAN	Mean	1.40	1.00	2.80	1.40	2.40	3.20	4.20	4.20
	N	5	5	5	5	5	5	5	5
	Std. Deviation	.548	.000	1.643	.894	.894	1.483	.447	.837
BN	Mean	2.40	3.40	5.00	3.20	3.20	4.80	5.00	4.00
	N	5	5	5	5	5	5	5	5
	Std. Deviation	.894	.548	.000	1.095	1.304	.447	.000	.000
LN	Mean	1.00	2.80	3.00	2.20	3.60	3.60	4.80	4.00
	N	5	5	5	5	5	5	5	5
	Std. Deviation	.000	1.095	1.581	.837	.548	.548	.447	.000
EAD	Mean	2.60	1.40	3.00	1.20	3.40	3.80	4.20	4.00
	N	5	5	5	5	5	5	5	5
	Std. Deviation	.894	.894	.707	.447	1.140	.837	.837	.707
LD	Mean	3.40	2.80	3.40	3.20	4.00	4.40	4.80	3.80
	N	5	5	5	5	5	5	5	5
	Std. Deviation	.894	.447	.548	.837	.000	.548	.447	.447
DD	Mean	1.60	1.20	3.60	1.20	3.20	3.00	4.60	2.80
	N	5	5	5	5	5	5	5	5
	Std. Deviation	.548	.447	1.517	.447	.447	.707	.548	.837
AHP	Mean	1.00	1.60	3.60	1.40	3.40	3.80	5.00	3.60
	N	5	5	5	5	5	5	5	5
	Std. Deviation	.000	.548	.548	.548	.894	.447	.000	.548
RO-MD	Mean	1.14	1.14	3.00	2.71	3.14	3.86	4.57	4.29
	N	7	7	7	7	7	7	7	7
	Std. Deviation	.378	.378	1.528	1.254	1.069	.378	.535	.756
Total	Mean	1.63	1.90	3.21	2.18	3.10	3.68	4.68	3.82
	N	62	62	62	62	62	62	62	62
	Std. Deviation	.910	.987	1.243	1.109	1.020	.937	.505	.713

Table 15: Mean scores for each writing exemplar for each panel.

Category		Script1	Script2	Script3	Script4	Script5	Script6	Script7	Script8
patients	Mean	1.30	1.95	2.80	2.35	2.70	3.40	4.75	3.75
	N	20	20	20	20	20	20	20	20
	Std. Deviation	.571	.887	1.105	1.040	1.129	1.046	.444	.716
nurses	Mean	1.60	2.40	3.60	2.27	3.07	3.87	4.67	4.07
	N	15	15	15	15	15	15	15	15
	Std. Deviation	.828	1.242	1.595	1.163	1.033	1.125	.488	.458
doctors	Mean	2.53	1.80	3.33	1.87	3.53	3.73	4.53	3.53
	N	15	15	15	15	15	15	15	15
	Std. Deviation	1.060	.941	.976	1.125	.743	.884	.640	.834
AHP	Mean	1.00	1.60	3.60	1.40	3.40	3.80	5.00	3.60
	N	5	5	5	5	5	5	5	5
	Std. Deviation	.000	.548	.548	.548	.894	.447	.000	.548
RO-MD	Mean	1.14	1.14	3.00	2.71	3.14	3.86	4.57	4.29
	N	7	7	7	7	7	7	7	7
	Std. Deviation	.378	.378	1.528	1.254	1.069	.378	.535	.756
Total	Mean	1.63	1.90	3.21	2.18	3.10	3.68	4.68	3.82
	N	62	62	62	62	62	62	62	62
	Std. Deviation	.910	.987	1.243	1.109	1.020	.937	.505	.713

Table 16: Mean scoring for each writing exemplar for each category.

Category	Subjective judgment Band	Score equivalent Band	Difference
Patients	Band 7.5	Band 8.5	+ 1
Nurses	Band 7.5	Band 7	- .5
Doctors	Band 7.5	Band 8.5	+ 1
AHPs	Band 7.5	Band 8.5	+ 1
RO/MDs	Band 7.5	Band 7	- .5

Table 17: Comparison of original subjective judgments and subsequent score equivalent bands for Writing.

4.3.2. Speaking

As described in Section 3.2.5.l)., above, each speech exemplar was ranked and subsequently allocated a score as follows:

5 = very good, better than acceptable (✓✓)
4 = acceptable speech for a doctor (✓)
3 = borderline acceptable (✓?)
2 = borderline not acceptable (x?)
1 = not acceptable speech for a doctor (x)

Table 18 shows the overall mean for each speech sample. As can be seen from Table 18, only four candidates are acceptable overall: HA and SA, representing a band-score level of Band 8 and AA and IK, both representing band-score levels of Band 8.5.

Table 19 shows the overall mean scores for each speech sample for each panel. As can be seen from Table 19, the acceptability of each candidate differs between panels although only 4 candidates are still acceptable overall. The speech sample from HN, representing a band-score level of Band 8, is acceptable to 10 panels and acceptable overall. Speech samples from AA and IK, representing band-score levels of Band 8.5, are acceptable to all panels, and overall. The speech sample from SA, representing a band-score level of Band 8, is acceptable to 6 panels and overall. However, the speech sample from BB, representing a band-score level of Band 7.5, although acceptable to 6 panels, is not acceptable overall. The speech sample from SH, representing a band-score level of Band 7 is acceptable to 4 panels, but not overall. None of the other speech samples is acceptable to any of the panels.

Table 20 shows the overall mean scores for each speech sample for each category. As can be seen from Table 20, only HA, AA and IK are acceptable to all categories but although SA is not acceptable to one category, she is acceptable overall. When comparisons are made with their initial judgments (Table 21), all categories except doctors make exactly the same subjective judgments as reflected in their score equivalents; doctors, on the other hand, give a band score equivalent one whole band higher than their mean subjective judgment.

	BB	NN	HA	AG	SA	RN	HN	AA	WL	SH	BR	IK
Mean	3.55	2.44	4.90	1.79	4.06	2.52	1.50	4.84	1.29	3.31	1.37	4.94
N	62	62	62	62	62	62	62	62	62	62	62	62
Std. Deviation	.918	1.065	.349	.926	.866	1.020	.763	.451	.611	.759	.579	.248

Table 18: Overall mean score for each speech sample.

Panel		BB	NN	HA	AG	SA	RN	HN	AA	WL	SH	BR	IK
LP	Mean	3.33	1.00	5.00	1.00	3.50	2.00	1.00	4.67	1.00	2.67	1.00	5.00
	N	6	6	6	6	6	6	6	6	6	6	6	6
	Std. Deviation	.516	.000	.000	.000	.548	.894	.000	.816	.000	.516	.000	.000
NP	Mean	2.71	2.00	5.00	1.71	3.71	2.71	1.43	5.00	1.14	3.29	1.14	5.00
	N	7	7	7	7	7	7	7	7	7	7	7	7
	Std. Deviation	.488	.577	.000	.756	.488	.756	.535	.000	.378	.756	.378	.000
CP	Mean	4.00	1.29	5.00	1.00	4.86	2.00	1.00	5.00	1.00	4.14	1.00	5.00
	N	7	7	7	7	7	7	7	7	7	7	7	7
	Std. Deviation	.577	.488	.000	.000	.378	.577	.000	.000	.000	.378	.000	.000
EAN	Mean	2.20	1.60	5.00	1.80	3.40	2.00	1.00	4.80	1.00	2.00	1.00	4.60
	N	5	5	5	5	5	5	5	5	5	5	5	5
	Std. Deviation	.447	.548	.000	1.095	.548	1.000	.000	.447	.000	.000	.000	.548
BN	Mean	5.00	4.00	5.00	3.20	4.00	3.20	2.80	5.00	1.60	4.00	2.20	5.00
	N	5	5	5	5	5	5	5	5	5	5	5	5
	Std. Deviation	.000	.000	.000	.447	.000	.447	.447	.000	.894	.000	.447	.000
LN	Mean	2.00	2.60	3.80	1.00	2.60	1.20	1.00	5.00	1.00	3.00	1.00	5.00
	N	5	5	5	5	5	5	5	5	5	5	5	5
	Std. Deviation	.707	.894	.447	.000	.894	.447	.000	.000	.000	.000	.000	.000
EAD	Mean	4.00	2.80	5.00	2.20	5.00	2.40	1.00	5.00	1.20	4.00	1.40	5.00
	N	5	5	5	5	5	5	5	5	5	5	5	5
	Std. Deviation	.000	.447	.000	.447	.000	1.140	.000	.000	.447	.000	.548	.000
LD	Mean	4.00	3.00	5.00	2.60	4.20	3.60	2.60	4.60	2.80	3.40	2.00	5.00
	N	5	5	5	5	5	5	5	5	5	5	5	5
	Std. Deviation	.000	.000	.000	.548	.447	.548	.548	.548	.447	.548	.000	.000
DD	Mean	3.80	2.60	5.00	1.00	5.00	1.80	1.00	4.60	1.20	4.00	1.40	5.00
	N	5	5	5	5	5	5	5	5	5	5	5	5
	Std. Deviation	.447	.548	.000	.000	.000	.837	.000	.894	.447	.000	.548	.000
AHP	Mean	4.00	2.40	5.00	1.40	3.40	2.60	1.20	5.00	1.00	2.60	1.00	5.00
	N	5	5	5	5	5	5	5	5	5	5	5	5
	Std. Deviation	.000	.894	.000	.548	.548	.548	.447	.000	.000	.548	.000	.000
RO-MD	Mean	4.00	3.86	5.00	2.86	4.71	3.86	2.43	4.57	1.43	3.14	2.00	4.71
	N	7	7	7	7	7	7	7	7	7	7	7	7

	Std. Deviation	.000	.378	.000	.690	.488	.378	.535	.535	.535	.378	.816	.488
Total	Mean	3.55	2.44	4.90	1.79	4.06	2.52	1.50	4.84	1.29	3.31	1.37	4.94
	N	62	62	62	62	62	62	62	62	62	62	62	62
	Std. Deviation	.918	1.065	.349	.926	.866	1.020	.763	.451	.611	.759	.579	.248

Table 19: Mean scores for each speech sample from each panel.

Category		BB	NN	HA	AG	SA	RN	HN	AA	WL	SH	BR	IK
patients	Mean	3.35	1.45	5.00	1.25	4.05	2.25	1.15	4.90	1.05	3.40	1.05	5.00
	N	20	20	20	20	20	20	20	20	20	20	20	20
	Std. Deviation	.745	.605	.000	.550	.759	.786	.366	.447	.224	.821	.224	.000
nurses	Mean	3.07	2.73	4.60	2.00	3.33	2.13	1.60	4.93	1.20	3.00	1.40	4.87
	N	15	15	15	15	15	15	15	15	15	15	15	15
	Std. Deviation	1.486	1.163	.632	1.134	.816	1.060	.910	.258	.561	.845	.632	.352
doctors	Mean	3.93	2.80	5.00	1.93	4.73	2.60	1.53	4.73	1.73	3.80	1.60	5.00
	N	15	15	15	15	15	15	15	15	15	15	15	15
	Std. Deviation	.258	.414	.000	.799	.458	1.121	.834	.594	.884	.414	.507	.000
AHP	Mean	4.00	2.40	5.00	1.40	3.40	2.60	1.20	5.00	1.00	2.60	1.00	5.00
	N	5	5	5	5	5	5	5	5	5	5	5	5
	Std. Deviation	.000	.894	.000	.548	.548	.548	.447	.000	.000	.548	.000	.000
RO-MD	Mean	4.00	3.86	5.00	2.86	4.71	3.86	2.43	4.57	1.43	3.14	2.00	4.71
	N	7	7	7	7	7	7	7	7	7	7	7	7
	Std. Deviation	.000	.378	.000	.690	.488	.378	.535	.535	.535	.378	.816	.488
Total	Mean	3.55	2.44	4.90	1.79	4.06	2.52	1.50	4.84	1.29	3.31	1.37	4.94
	N	62	62	62	62	62	62	62	62	62	62	62	62
	Std. Deviation	.918	1.065	.349	.926	.866	1.020	.763	.451	.611	.759	.579	.248

Table 20: Mean scores for each speech sample from each category.

Category	Subjective judgment Band	Score equivalent Band	Difference
Patients	Band 8	Band 8	=
Nurses	Band 8	Band 8	=
Doctors	Band 7	Band 8	+ 1
AHPs	Band 7.5	Band 7.5	=
RO/MDs	Band 7.5	Band 7.5	=

Table 21: Comparison of original subjective judgments and subsequent score equivalent bands for Speaking.

4.4. Confirmatory analysis of the Speaking and Writing Judgments

4.4.1. A brief overview of Many-Facet Rasch measurement

In order to confirm the findings of the panels for the speaking and writing judgments, the data for all participants were analysed using a procedure known as many-facet Rasch (MFR) analysis using the programme FACETS Version 3.70.0.

Many-facet (also referred in the literature as *multi-faceted*) Rasch analysis is a variation on a logistic regression model developed by the Danish mathematician Georg Rasch (1960). Each observation or judgment is considered as being influenced by a number of factors (or facets), for example the judge (harshness, consistency, etc.), the participant (ability to perform the assessed task), the task being undertaken (difficulty). The different facets contribute independently in the model, to a probabilistic estimate of the true value or level of the object or individual being judged: this might relate to a person's ability to communicate effectively in a medical setting as judged by experts or peers, or it might refer to the judgments made by panel members in a standard-setting event.

The standard output from a many-facet Rasch (MFR) consists of a series of tables representing probability estimates for each facet included in the analysis (so if there are three facets there will be three tables). The columns contained in these tables (see for example Table 21) are explained below:

Total Score This is the sum of all of the judges' observations (i.e. 1 = not acceptable; 5 = very good)

Total Count This is the total number of judgments made (i.e. in this case there were 57 judges whose observations were included in the analysis.)

Observed Average This is the raw average (i.e. Total Score divided by Total Count.)

Fair-M Average This is the Average score when the various facets (i.e. the judges, the task of judging and the application of the five point scale) are taken into account. This is an estimate of the true level of acceptability of each testee (in this case the

	examples of test performance) reported on the five point scale.
Measure	This is actually then same as the previous column, but the result is reported on a logit (pronounced *loh-jit*) scale. This is useful for measurement purposes as all facets are reported on this same scale, which can be compared across the different facet tables.
Model Standard Error	In classical test data analysis, we estimate a single standard error (which is used as an indication of the accuracy of the test). Facets improves on this approach by creating a standard error for each element of each facet
Infit & Outfit Mean Square	Infit and Outfit are indicators of how accurately or predictably data fit the model and are sensitive to different types of responses. The expected value for both is 1.0 – below 1.0 means the judgments are too predictable and above 1.0 means that the judgments are unpredictable. Linacre (2002: 878) suggests that estimates above 2.0 are likely to be problematic, while 0.0 to 0.5 and 1.5 to 2.0 are less useful but unlikely to be seriously concerning. The range 0.5 to 1.5 is considered to be the most productive – though as explained in table 20, this range will vary depending on the type of judgments made.
Infit Mean Square	Infit reflects inlier-sensitivity: e.g. where a judge is idiosyncratic in his/her judgments. Due to their nature, these patterns are difficult to interpret. For this reason, Infit scores are usually reviewed first and where they are problematic are used as the basis of any decision to include or remove a judge from an analysis. The judges removed from the MFR analysis all displayed higher than expected Infit mean square statistics.
ZStd	These are the standardised fit statistics and are actually a t-test of the hypothesis "Do these data fit the model perfectly?" An estimate below 0.0 reflects predictability while one above 0.0 reflects unpredictability. Interpretation will depend on the focus of the analysis – so, in this case a highly predictable result reflects the high level of agreement among the judges (with harsh ones always marking it down and lenient ones always marking it up). In this case, we are not concerned with negative estimates, but would note that high positive estimates reflect the difficult encountered by judges in making their decisions.

Outfit Mean Square	Outfit mean square reflects outlier sensitivity: e.g. they may reflect carelessness in judgment making. These are considered to be less problematic (in terms of measurement) than very high Infit mean square estimates.
Estimated Discrimination	As the title suggests, this is an estimate of the discrimination of each element of the facet. The expected value is 1.0 and the range 0.5 to 1.5 is considered to indicate reasonable fit to the model (Linacre, 2012:185)
Correlation	The two correlation columns represent the observed and the expected correlations. Where the data fit the model, we expect that these will be similar so, in Table 21, the trend across all 12 elements is very similar and therefore these correlations would be considered acceptable.
Correlation PtMea	The point-measure correlation is the correlation between the observations and the measures modelled to generate them when the data fit the Rasch model" (Linacre, 2012 :185)
Correlation PtExp	The expected values of the point-measure correlation is the expected correlation between the observations and the measures modelled to generate them" (Linacre, 2012 :185)
Nu	This is the number of the elements in the analysis
Testees	This is the name given to the elements (e.g. 1 = BB)

The application of MFR has become quite common in the area of educational assessment; McNamara & Knoch (2012) provide a useful overview of its use in the area of language testing and it has also been applied in the area of medical research. In the medical domain, the areas researched have tended to focus on the development of a judgment-based instrument (e.g. Beck & Gable, 2000, 2003; Liao & Campbell, 2002; Ibey et al., 2011), research based on patient or expert judgment (e.g. Lai et al., 1997; Darragh et al. 1998; Ahlstrum et al., 2004; Fitzpatrick, et al., 2003; Hayase et al., 2004; Decruynaere, et al., 2007) or on providing an introduction to the use of MFR in a particular area of medical research (e.g. Pallant & Tennant, 2007). However, the most common use of MFR has been in validating existing tests or instruments in which judgments are made (e.g. Bernspang & Fisher, 1995a, 1995b; Atchison et al., 1998; Bernspang, 1998; Girard et al., 1999; Beck & Gable, 2001a, 2001b; Campbell et al., 2002; Bode et al., 2003; Kottorp et al., 2003; Liao & Campbell, 2004; Malec, 2004; McManus et al., 2006; de Morton et al., 2008; Chien et al., 2012).

4.4.2. The MFR analysis of the Speaking judgments

The results of the initial MFR analysis indicated that five judges (20, 32, 34, 46, 56) were somewhat inconsistent in their judgments and, as a result, these individuals were removed from the next run of the analysis. The five judges removed from the MFR analysis had all displayed higher than expected Infit mean square statistics; on removing them, the final run indicated that the remaining judges tended to be consistent, though varying in terms of harshness/leniency.

According to Lunz and Wright (1997: 83) "Because the interpretation of fit is situationally dependent, there are no fixed levels for fit statistic acceptance or rejection." Wright and Linacre (1994) suggest the range of acceptable figures shown in Table 22. While the task undertaken in this project is certainly not a clinical observation, the situation where the individual judges are expected to offer a range of experience and expertise, is similar in measurement terms in that we are neither expecting nor encouraging agreement among the judges. For this reason, we opted to allow for a broader range of Infit and Outfit mean square statistics than recommended by Wright and Linacre (1994:370).

Type of Test	**Range**
MCQ (High stakes)	0.8 - 1.2
MCQ (Run of the mill)	0.7 - 1.3
Rating scale (survey)	0.6 - 1.4
Clinical observation	0.5 - 1.7
Judged (agreement encouraged)	0.4 - 1.2

Source: Wright & Linacre (1994:370)

Table 22: Reasonable Item Mean-square Ranges for INFIT and OUTFIT

The consistency is estimated by looking to the Infit and Outfit Mean Square columns on the output (Table 7.1.1. Judges' Measurement Report in Appendix 5a). From this column it is clear that EAN 1 (judge 21) and EAN 3 (judge 23) are somewhat inconsistent as their Infit and Outfit Mean Square statistics are higher than usually recommended. We considered removing these from the analysis but found no improvement to the model when this was done. All others were found to be acceptable. This is not a particularly surprising finding given that the individual participants had a clear idea of the purpose of the event and would have come to it with a keen sense of the language ability they would expect a doctor to exhibit. They were also offered an opportunity to discuss the notion of a minimal level of language competence with the other judges on their panel and were demonstrated the test and the judging task before they made any decisions.

The Testee Measurement Report (Table 23) is reproduced from the full output and is an indication of the cumulative judgment of all participating panel members on the appropriateness of the test performances they were asked to judge.

The Infit and Outfit Mean Square columns suggest that, with the possible exception of performance AA, panel members found it quite easy to place the candidates on the scale. The slightly high numbers for AA's performance suggest that there was an unexpected disagreement amongst judges as to the acceptability of the speaker.

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Nu	Testees
204	57	3.58	3.64	-.50	.22	.78	-1.1	.75	-1.3	1.26	.76	.66	1	BB
134	56	2.39	2.41	2.18	.20	.79	-1.2	.82	-1.0	1.25	.81	.73	2	NN
282	57	4.95	4.98	-6.94	.61	.88	.0	.30	-.2	1.11	.34	.25	3	HA
103	57	1.81	1.66	3.53	.21	.81	-.9	.89	-.3	1.23	.75	.70	4	AG
233	57	4.09	4.11	-2.00	.24	1.35	1.6	1.46	2.1	.52	.52	.64	5	SA
148	57	2.60	2.65	1.76	.19	.96	-.1	.93	-.3	1.05	.72	.73	6	RN
86	57	1.51	1.33	4.40	.24	.71	-1.4	.64	-1.0	1.29	.73	.63	7	HN
277	57	4.86	4.92	-5.73	.41	1.69	2.1	6.60	3.8	.13	-.08	.39	8	AA
74	57	1.30	1.16	5.22	.29	1.08	.3	.63	-.6	1.13	.58	.54	9	WL
187	57	3.28	3.35	.24	.20	.97	.0	1.02	.1	.98	.54	.68	10	SH
80	57	1.40	1.24	4.78	.26	.47	-2.7	.45	-1.5	1.34	.75	.59	11	BB
282	57	4.95	4.98	-6.94	.61	.91	.0	2.09	1.0	.98	.22	.25	12	IK
174.2	56.9	3.06	3.04	.00	.31	.95	-.3	1.38	.0		.55		Mean (Count: 12)	
77.7	.3	1.36	1.44	4.31	.15	.30	1.3	1.64	1.5		.26		S.D. (Population)	
81.1	.3	1.42	1.50	4.51	.15	.31	1.3	1.71	1.6		.27		S.D. (Sample)	

Model, Populn: RMSE .34 Adj (True) S.D. 4.30 Separation 12.62 Strata 17.16 Reliability .99
Model, Sample: RMSE .34 Adj (True) S.D. 4.49 Separation 13.18 Strata 17.91 Reliability .99
Model, Fixed (all same) chi-square: 1601.1 d.f.: 11 significance (probability): .00
Model, Random (normal) chi-square: 10.9 d.f.: 10 significance (probability): .37

Table 23: Testee Measurement Report Speaking

The Observed Average column shows the mathematical mean of the judgments made, while the Fair Average column indicates the probability of the true value of this mean when other factors (rater harshness, application of the scale) are taken into account. This column indicates that just four of the performances are deemed acceptable (HA, SA, AA and IK), supporting the findings of the qualitative and quantitative results reported above.

4.4.3. The MFR analysis of the Writing judgments

As was the case with the speaking judgments, initial MFR analysis indicated that eight judges (1, 9. 25, 26, 36, 37, 49, 61) were somewhat inconsistent in their application of the basic scale devised for this project (a 5-point acceptability scale where 4 was the minimum acceptable level). When these individuals were removed from the analysis the resulting output was found to be more stable, though a number of individuals still remained inconsistent to some degree. As explained in 4.4.2. we opted to allow for a broader range of Infit and Outfit mean square statistics than recommended by Wright and Linacre (1994:370); the range of acceptability for judges was therefore considered to be acceptable up to 2.0 on the basis that we were neither expecting nor encouraging total agreement among the judges. All judges included in this analysis fall into this range.

The results of the final MFR analysis of the writing data indicate that just one sample was adjudged by the participants (Script 7) to have reached the level of writing ability

required of a practicing doctor. This again supports the findings of the previous analyses (see Table 24).

The Infit and Outfit columns suggest that the judges did not find these scripts very difficult to judge, with the slight exception of Script 3. It is clear from the subjective data that writing (along with reading) was the most contentious of the areas, with a broader range of performances seen as being adequate. Despite this, when we look at the recommendations of the panels from the subjective data, writing stands out as being the only area where consensus was found among all stakeholder groups. However, the instability implied by the MFR output reflects the fact that the standard deviation for this performance is considerably greater than for the other scripts (see table 14 above). This suggests that there was a significantly broader range of opinions about the acceptability (or not) of this script than there was for any of the others.

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	N	Candidates
87	54	1.61	1.47	2.36	.20	1.00	.0	1.05	.2	1.00	.59	.61	1	1
99	54	1.83	1.72	1.94	.18	.71	-1.5	.71	-1.2	1.13	.69	.63	2	2
178	54	3.30	3.39	-.09	.17	1.48	2.2	1.54	2.4	.40	.50	.54	3	3
117	54	2.17	2.13	1.43	.16	.86	-.7	.91	-.4	1.19	.66	.63	4	4
169	54	3.13	3.23	.15	.16	.65	-2.1	.63	-2.1	1.44	.62	.56	5	5
200	54	3.70	3.76	-.79	.19	.89	-.4	.85	-.6	1.05	.66	.50	6	6
256	54	4.74	4.78	-3.91	.31	.97	.0	.85	-.4	1.04	.28	.31	7	7
208	54	3.85	3.89	-1.10	.20	.92	-.3	.86	-.6	1.04	.23	.49	8	8
164.3	54.0	3.04	3.05	.00	.20	.94	-.4	.93	-.4		.53		Mean (Count: 8)	
55.1	.0	1.02	1.09	1.89	.05	.24	1.2	.26	1.3		.17		S.D. (Population)	
58.9	.0	1.09	1.16	2.02	.05	.25	1.3	.28	1.3		.18		S.D. (Sample)	

Model, Popuin: RMSE .20 Adj (True) S.D. 1.88 Separation 9.31 Strata 12.75 Reliability .99
Model, Sample: RMSE .20 Adj (True) S.D. 2.01 Separation 9.96 Strata 13.62 Reliability .99
Model, Fixed (all same) chi-square: 511.3 d.f.: 7 significance (probability): .00
Model, Random (normal) chi-square: 6.9 d.f.: 6 significance (probability): .33

Table 24: Testee Measurement Report Writing

It is interesting that the fair average scores awarded for the speaking (3.04) and the writing (3.05) are almost identical. Since the performances provided to the participants represented a broad range of band scores, this suggests that the participant behaviour during both parts of the judging process was similar, i.e. they did not expect more of one skill over the other.

It is also interesting that the results of the analysis of the writing scores reported in this section, in keeping with the analysis of the speaking scores reported in 4.3.1., fully support the findings from the other analyses. Whilst not wishing to claim that one type of analysis validates another, it is clear that the MFR analysis adds considerable weight to the findings of the qualitative and quantitative results reported above in 4.1 and 4.3, thus offering a significant degree of triangulation to the findings and strengthening the basis on which the recommendations that follow in Section 5 are based.

4.4. Comments from initial stakeholder panels on the IMG-EEA distinction

As mentioned in Section 3.2.5.(3), the current situation regarding IMGs and those from the EEA or holding EC rights was explained to each panel. Each panel then discussed the IMG/EEA distinction issue to decide if both should be treated as a single population with the same language requirements or as separate populations with different language requirements (Question 3).

Each panel was asked the following question:

> If evidence of English language competence should ever be required of EEAs in order for them to be admitted to the GMC register, should they provide the same evidence as IMGs?

In response to this question, every panel gave the following response (or a very close variant of it – see Appendix 3):

N *Yes, they should all provide the same evidence.*

All comments on this issue can be found in Appendix 3. Comments can be divided into three main themes: comments relating directly to the question, above; comments relating to broader issues arising from it including the testing of UK graduates' language and communication skills; comments relating to the main purpose of testing English language skills, namely, that of ensuring patients' safety.

Comments addressing the specific question included:

N. *I think they should all be the same.*

N *Anybody who is not a native speaker of English should provide the same standard of evidence of language ability.*

N *We should ask EEA doctors for the same evidence of language ability that we ask of IMG doctors.*

D *I think anyone whose first language is not the language they are going to be conducting their job in should have to exhibit a proficiency in that second language.*

P *Everybody should provide the same evidence of language ability, it's ridiculous not to.*

D *It's only reasonable that both IMG and EEA, rest of the world and Europeans should all provide the same evidence of ability to speak English.*

AHP *I think everybody should be made to provide evidence, I don't see how you can differentiate in any way.*

RO *If there is a sort of escape route for IMGs to become EEA then that is an issue so all IMG and EEA graduates should provide the same evidence of language ability.*

D *Just because employment law in the EEA means that people can move freely and work freely doesn't lessen the language barriers that they are going to face so I think I agree there should be the same requirements.*

A few panel members were in favour of possibly allowing exemptions from a language test:

AHP *I think if you have done your degree in English as your first language then perhaps you can have that waived but not if you have done university in another country.*

RO *The issue is more about which language you are educated in isn't it. If you are educated in English even in France you could argue that the person's language skills ought to be adequate.*

Several of the comments relating to broader issues arising from the responses to the question, particularly from the RO/MD panel, concern the issue of also including UK graduates:

RO *There should be no difference between a graduate from the UK or an IMG or a EEA doctor.*

RO *There is a separate argument about whether UK graduates should be subjected to the same test.*

RO *You can come in as a member of your family and become an undergraduate and not have the language skills. That is an entry to your undergraduate career. We should start out with that base. I think it is important that the issue about UK graduates is also considered.*

RO *If there is a principle of fairness then the argument to test everyone is the right one to take.*

RO *They should decide to put English or language testing into medical school curriculum; that would mean you couldn't graduate without achieving a standard of communication.*

D *This is fundamentally about patient safety so despite the extra cost it might incur really everyone should undergo a test and they should go through the same rigorous test.*

The issue of ensuring patient safety was further elaborated:

RO *Our first duty has to be to the patients.*

D *The bottom line is patients' safety and we can't ensure patients' safety unless we ensure communication.*

D *The fact that you have a medical degree doesn't mean to say that you are a safe doctor, part of being a safe doctor is being able to communicate.*

P *It's only fair if they come to work in this country that people should be able to communicate in the language of the country.*

RO *If reception of what the patients are saying is misinterpreted it can go in all sorts of directions that aren't appropriate.*

RO *Consequences are much greater than going to the supermarket and ending up with the wrong vegetables.*

P *It's a very serious issue, it's life or death.*

4.5. Summary of results from initial stakeholders' panels

The research was undertaken in order to answer the questions which arose from the literature review.

Question 1:

Does the IELTS offer an appropriate measure of the language ability of prospective medical professionals?

It would seem that the answer to this question is far from clear cut. From the comments presented in Section 4.1., it is apparent that most of the panels thought that the IELTS is a reasonably good test of overall English language ability. However there is also a clearly articulated belief that IELTS is not sufficient on its own without some form of further test which is more medically-oriented. This is reflected in the following comments (See Appendix 2 for comments relating to each IELTS skill and to IELTS overall, many of which are very positive about the IELTS test as a general English language test):

N *I think the test is limited.*

N *If you are saying that this is a test that we are going to use to see whether an individual has the communication skills in whatever format to function as a member of a medical team, then you could get much more useful information than from that. This is way too broad.*

N *It's a good indicator of whether somebody has sufficient English to come to this country and apply for the PLAB test.*

N *I think it's a starting point.*

P *I think there should be something more. I think we all feel there ought to be additional tests.*

P *I mean, these test are not adequate.*

P *There should also be an integrated test or possibly some additional tasks to determine whether doctors could perform the sorts of things that doctors need to do in each of the skills.*

D *I think this is a screening test.*

D *I think you won't be able to test the medical side of things unless you actually start testing specifically the 'can do' statements. Because at the start we said that we think these are the best set we've had so if you design a test to address those statements that might give you a better idea of someone's capability for functioning as a doctor.*

D *It's almost a focusing test to see whether they should then proceed to the real doctoring test.*

D *In my opinion there needs to be another test, or the PLAB needs to be changed and have a better focus on it, because I don't think this is going to give you enough information as it is about how someone is as a doctor.*

AHP *I think they could stand to do this test of their understanding and comprehension and general English skills but that there should also be an additional, career specific, test.*

Most panels were favourably impressed with the reading and speaking tests, less so with the writing test, but the listening test was viewed by all as particularly problematic as a test of listening skills for doctors.

N *They speak very clearly, the pronunciation is very, very clear and it runs chronologically so if we are going to assess somebody's ability to listen to real people in real life, I don't think it's a particularly useful tool to assess people against what we are actually expecting them to do.*

N *It's all actors makes it very stilted. There's no variety really in what we are asking them to do and the variety of listening they are going to have to do in their jobs isn't reflected in this.*

N *You could do very well on this test but how you function in an environment hasn't been tested at all.*

P *I don't think it reflects what they are going to hear either so if they can't get the simple stuff on these tests they have no hope in a busy hospital really.*

P *No way does it test the things we said earlier that doctors have to listen to.*

P *This test won't give an indication of them (doctors) being able to do the things we think they should do.*

P *I don't think these tests are relevant to listening skills that a doctor would need. I think these are fine for somebody who wanted to come here to study but not in a busy situation that has background noise.*

P *There certainly needs to be more diversity in the kinds of skills they are testing.*

P *It's an unnatural test in a way that it doesn't represent what they will meet when they come here.*

D *I think the one thing it misses out on is understanding whether a person has got the whole gist of the situation.*

D *This audio is a lot easier to understand than a lot of hospitals, in a busy A&E and regional accents and colloquialisms, none of that. And people don't speak in such neat sentences.*

D *All this test seems to be doing is writing down what you have heard.*

D *I think part of the trouble is we are trying to assess each part of the language separately whereas they are all interlinked and you can't assess the full complexity of one aspect of it without assessing the other components at the same time.*

D *It's a listening test but it's not at all appropriate as a measure for assessing whether a doctor can cope in a clinical setting of any description.*

D *I don't think this test is appropriate at all and none of the skills are really tested.*

D *The listening test is just farcical.*

AHP *I found it to be forced, unnatural conversation. It wasn't like natural conversation where it was flowing very quickly and you had to pick things up fast.*

AHP *It needs something with background noise or distractions.*

AHP *It's a good test of general listening ability but it's too clean, it's not listening in the real world and some of it is too artificial and structured.*

AHP *Doctors need to be able to pick up information from much more messy discourse with background noise.*

RO *If they can't do this test, they won't be able to manage in A&E where they've got anxious relatives yelling at them and patients who are pretty incoherent.*

RO *They need to listen over the phone. Listening face-to-face and listening over the telephone are two important and sometimes different methods of communication.*

RO *They need a test like the call handlers 999 ambulance services…. with agitated, stressed people.*

From the above comments it would seem that some of the skills deemed essential for doctors (See Appendix 1 for the 'can do' statements) that they feel are not adequately tested in the IELTS listening test are:

Can listen and pick out relevant and important information from patients: adults/children/elderly/disabled/brain damaged/otherwise impaired/drunk

Can pick out relevant and important information in emergency situations while doing other things/ in excitable or stressful situations with considerable background noise

Can follow extended, natural speech even when it is not clearly structured/ extract the gist of what is said when conversation is animated and delivered at a fast natural rate in a range of accents/ extract gist, detail, purpose and main points from formal discussions involving detailed information such as facts or definitions related to professional topics/ integrate information from multiple sources and follow detailed instructions to carry out complex tasks involving unfamiliar process or procedures.

Notwithstanding the limitations concerning various aspects of the IELTS test, there is sufficient evidence from the comments about the separate skills and the test overall to suggest that, given the current lack of a specific, medically-oriented English language test, it is worth retaining the IELTS test as an instrument to assess general English language proficiency.

Question 2:

If it is found that IELTS is an appropriate instrument, what is the most appropriate level for prospective medical doctors to attain before being accepted to practise in the UK?

It would seem from the results presented in this section that there are no clear cut answers to this question, either. Subjective judgments averaged for all categories of panels suggest that Band 7.5 overall, with no skill lower than Band 7.5, but with listening at Band 8.5, are the preferred levels. However, the subjective judgments of both patients and nurses (representing 35 panel members or 56% of the total participants) were that the minimum requirement should be Band 8 overall, with a level of Band 8.5 required for listening but Band 7.5 acceptable for writing. In fact, most panels found it very difficult to agree on an overall Band score and were reluctant to specify anything other than a profile. In the end, most agreed to an overall Band score by averaging the skills requirements.

Descriptive statistics analysing the scores for reading and listening and the scores awarded to writing and speech exemplars present a different profile of a minimally competent candidate as follows: Reading = Band 8; Writing = Band 8.5; Listening = Band 8.5; Speaking = Band 8. MFR analysis of the writing and speaking scores confirms the findings of the descriptive, quantitative analysis.

Using the IELTS overall Band score system, these scores would be averaged to produce an overall requirement of Band 8 or possibly even Band 8.5 as a mean score of 8.25 would be rounded up to 8.5. However, although it is certain from the comments made about the listening test that all panels would insist on the requirement for listening being recorded as Band 8.5, it is far less clear that any of the panels would agree to the writing requirement being set at 8.5.

Subjective judgments, descriptive analysis of score-equivalent bands awarded on each IELTS skill and MFR analysis of the writing and speaking skills are presented in Table 25. .

Skill	Subjective judgment Band	Score equivalent Band	MFR analysis
Reading	Band 7.5	Band 8	n.a.
Writing	Band 7.5	Band 8.5	8.5
Listening	Band 8.5	Band 8.5	n.a.
Speaking	Band 7.5	Band 8	8

Table 25: Subjective judgments, band score equivalents and MFR analyses of the IELTS skills

In addition to other facets affecting the complexity of assessing speaking performances previously mentioned, the difficulty of getting raters' agreement on writing tests is also well-documented (cf. Elder et al., 2005; Elder et al., 2007; Knoch, 2009; Knoch et al., 2007; O'Sullivan & Rignall, 2007; Stahl & Lunz, 1996; Weigle, 1998, 1999; Wigglesworth, 1993, for discussions of aspects of rater training and rater feedback in an attempt to increase inter-rater reliability). However increasing rater-reliability may in itself be problematic as Eckes (2009:5) points out:

> *Trying to maximize interrater reliability may actually lead to lowering the validity of the ratings, as would be the case, for example, when raters settled for attending to superficial features of examinee performance (see Hamp-Lyons, 2007; Reed & Cohen, 2001; Shohamy, 1995). This clearly unwanted effect is reminiscent of the attenuation paradox in classical test theory (Linacre, 1996; Loevinger, 1954).*

It was pointed out earlier that the previous study in 2004 had encountered similar problems in deciding on an appropriate writing cut score as each of the three stakeholder groups set their cut-off at a different point within the range of Band 5 – Band 8. If we revisit Table 16, we see that in the current study, only script 7 (Band 8.5) reaches (and exceeds) the cut-off of 4; however it could be plausibly argued that script 8 (Band 7) might be rounded up from 3.82 to 4 making Band 7 the cut score. It could even be argued at a stretch that Script 6 (Band 6) could be rounded up from 3.68 to 4 making the cut score Band 6. This of course is confounded by the problem of Scripts 3 (Band 7.5) and Script 5 (Band 8) not reaching a sufficiently high average score to be reasonably rounded up to 4 and therefore being rejected.

Commenting on the writing cut scores to be set in the 2004 study, Banerjee (2004: A7) comments: "The GMC will need to decide how best to reach a compromise between these positions". Since we prefer to provide more explicit recommendations, if possible, all results were presented to the final panel for discussion and ultimately for a decision on the appropriate cut scores for each skill and overall.

Question 3:

> *Should all non-native-English speaking doctors be asked to empirically demonstrate their language level?*

Or, to rephrase this question more specifically as it was put to the initial stakeholder panels:

> *If evidence of English language competence should ever be required of EEAs in order for them to be admitted to the GMC register, should they provide the same evidence as IMGs?*

From the comments presented above in relation to this question, there is no doubt that it can be answered without equivocation. Every member of every panel stated explicitly that all non-native speakers of English should provide evidence of their English language ability regardless of nationality or origin. It was also suggested by one panel that all applicants to the GMC register should provide evidence of their English language ability, including those of British nationality. Participants on the patients' panel in Caernafon in North Wales, where there is a large population who speak Welsh as a first language, also commented that it was important for any doctor wishing to practice there to have at least a willingness to learn Welsh.

5. Final panel deliberations and recommendations

5.1. Final Panel

Following completion of the initial stakeholders' panels, all judgments and data were analysed and comments on each of the IELTS skills tests and overall impressions of the test were collated. Representatives from each category of panel were invited to participate in a final confirmatory panel in London on 12 October to discuss the judgments and quantitative results derived from the initial panels and make recommendations regarding the IELTS test for the GMC.

5.1.1. Final panel composition

The final panel comprised 8 members drawn from the original participants of the initial stakeholders' panels; two patients (one male, one female), two nurses (one male, one female), two doctors (two males), one allied health professional (female) and one responsible officer (male). Participants came from different geographic locations of the UK and were reimbursed for travel expenses, provided with hotel accommodation where necessary, and paid an incentive of £310 for the day's work.

5.1.2. Final panel procedures

The final panel was organised as follows:

1. The convener reminded the panel of the basic aims and objectives of the research and reviewed the methodology of the study.
2. Participants were presented with summaries of 'can do' statements for all skills (see Appendix 1) and summarized comments from all the initial panels on each of the IELTS skills tests and overall (see Appendix 2).
3. Summaries of qualitative judgments for each skill were presented by panel (individual panels' judgments) and by category of panels (patients', doctors', nurses' average judgments) and discussed; this discussion, and all subsequent discussions, was recorded for later transcription and review.
4. Quantitative analyses were presented (descriptive statistics, including means and ranges of scores for Reading and Listening, converted to Band scores, plus means of scores ranging from 1 – 5 which had been allocated to speech samples and writing exemplars) and discussed.
5. Required IELTS scores for registration of overseas doctors by medical councils in Australia, Canada, Ireland, New Zealand and South Africa were presented as were requirements for registration with non-medical professional bodies within the UK (see Appendix 6a and 6b) and discussed.
6. Decisions were made regarding appropriate levels to recommend for each skill and overall.
7. The appropriacy of the IELTS as a screening test for overseas doctors prior to registration for PLAB Part 1 was discussed and decisions on recommendations were made.

8. Language requirements of EEA countries for overseas registration of doctors were presented and discussed.
9. The current IMG-EEA distinction in the UK was discussed and decisions on recommendations were made (see Appendix 3).
10. Overall recommendations to be offered to the GMC were noted, related back to the final panel, further discussed and confirmed.

5.1.3. Final panel discussion

Although the intended organisation of procedures for the final panel described in 5.1.2., above, is clearly structured and linear, in reality it was found to be impossible to follow to the letter. As all final panel members had already participated in an initial stakeholders' panel, they were very familiar with the format of the IELTS test and had formed their own opinions on the usefulness of it both as an overall proficiency test and of each of the skills areas separately. It was not possible to conduct the panel by discussing each of the skills areas separately as originally intended. Although all points in the procedures were covered, discussion of points presented in this section is not linear, but organised to address the various themes that emerged over the course of the day.

a) Representativeness of initial stakeholder panels

The panel discussed whether the 62 initial stakeholder panel participants could be considered to be representative of the population of the UK as a whole. Given the diversity of gender, age, ethnicity, work experience and geographic location achieved, it was felt that, despite the relatively small sample size, the balance was as good as could have been attained, given that the main criterion had simply been a willingness to take part. It was thought at the time that we may have had an over-representation of ethnic minority participants as the results of the 2011 census had not been published when the panels were convened; following publication of the census in late 2012, it appears that we have exactly the same percentage of ethnic minority participants as are currently resident in England and Wales, which accounts for over 92% of the population of the UK (data are not available for Scotland and Northern Ireland). The gender balance is also representative according to official census figures of just under 51% female and over 49% male[9].

b) Objective versus subjective legitimacy: judgments versus scores in the receptive skills tests

As described earlier in Section 2.4.2., because the IELTS test comprises two receptive skills papers (listening and reading) and two productive skills papers (speaking and writing), two different approaches to standard-setting are required in this study. For the receptive papers the most appropriate approach for the purposes of this study was identified as the modified Angoff method (Hambleton & Plake, 1995) with results rounded up to the nearest integer.

[9] http://www.ons.gov.uk/ons/rel/census/2011-census/population-and-household-estimates-for-the-united-kingdom/stb-2011-census--population-estimates-for-the-united-kingdom.html

In agreement with the principal of standard-setting identified for the receptive papers it was argued that the results of the initial panels should be based on the objective scores given rather than on the initial subjective decisions made. Using the reading scores as an example, it was pointed out that when the decisions on band scores are averaged, the overall requirement would be Band 7.5; but if the actual scores awarded are analysed, a mean calculated, and the mean subsequently converted to a band score, the overall requirement would be for Band 8 for reading. Comments relating to the calculation of the reading mean included:

FP4 *This is an assessment of scores that we actually gave. There is an argument, isn't there, for straightening the requirement to a Band 8 for reading? It's more objective.*

FP2 *Given that the scores are very close between the groups I think that is quite a telling point, so we don't have a wide range of figures. It means everybody was on the same kind of wave length. It doesn't seem wrong to expect a higher standard. My feeling now is that the general opinion is that we want a higher standard.*

In terms of the listening scores, this was not an issue as both methods of calculation produced the same band score requirement.

c) Patients' versus professionals' views

There was much discussion about the importance of patients' views and clear agreement with the following comment:

FP4 *Another justification for raising the scores rather than lowering them is taking the patients' view more seriously than the professionals' view. Because professionals can be expected to interpret what we talk to each other about and what we write and what we read so our communication is, we can interpret things better. But as a patient if you are struggling to cope with a doctor or a professional of any sort who doesn't speak your language very well you are at a much greater disadvantage than a professional is. So I would wish to be guided by the patients' view if it varies from my previous view.*

Other comments supporting this view included:

FP6 *I think that's a very good point. I mean especially as a patient, the ability to connect on a less than technical level and actually have a discussion that you can understand because it's your health that's paramount.*

FP8 *Taking the patients more seriously than the doctors' and nurses' views, I think that is a very good point.*

FP6 *And I think it's important to be able to think that you can challenge a medical professional, not in an aggressive way but about alternative treatments or any sort of questions that you might have. I think it's important enough to be assured of a clear response.*

d) Decisions on IELTS band scores for each skill and overall

As mentioned in b) above, for the receptive papers (reading and listening) the most appropriate approach for the purposes of this study was identified as the modified Angoff method (Hambleton & Plake, 1995) with results rounded up to the nearest integer.

With regard to standard-setting for the performance-based language tests (writing and speaking), as explained in 2.4.2., the most appropriate method was identified as an examinee-centred approach, such as the 'Examinee Paper Selection' method (Hambleton, et al., 2000), also known as the Benchmark method (Faggen, 1994). In this approach, judges make decisions based on test task performance (recordings of speech or copies of written work) rather than by reviewing the task input material from the test paper. Although the original intention had been for the final panel to discuss each skill separately and arrive at an agreed Band score requirement for each skill and then overall, this proved to be impossible in practical terms for reasons explained earlier in the introductory paragraph to 5.1.3. This also may be a reflection of a comment from one of the initial stakeholder panels:

D *I think part of the trouble is we are trying to assess each part of the language separately whereas they are all interlinked and you can't assess the full complexity of one aspect of it without assessing the other components at the same time.*

Discussion of IELTS skills and decisions on appropriate Band scores will therefore be presented as discretely as possible but will be grouped together in terms of receptive skills (reading and listening) and productive skills (writing and speaking) as at times there is considerable overlap. Finally, comments on the overall Band score to be required will be offered. . (Participants are labeled as FP1 – FP8; Mod is the lead moderator.) For all comments made by the final panel on all IELTS skills and IELTS overall, please see Appendix 2.

Receptive skills – Reading and Listening

The discussion began with a general review of the test and the order in which the skills were presented to the initial stakeholder panels. One participant was concerned that the group score given by them for **reading** may have been skewed by affective factors rather than decisions about reading ability:

FP7 *I wonder looking at us, I think we probably got more confident in our decision making as the process went on and actually I would be inclined to think that our 7 for reading was probably a little on the lenient side, whereas towards the end our figures were much more 7.5, 8.5, so I wonder whether that skewed things slightly for our group in particular for the reading score.*

The participants then considered the content of the reading test and discussed how well it reflected the actual reading skills a doctor would need:

FP7 *I would say as well with the reading it's all well and good to think with reading, OK you have plenty of time, but actually, realistically, in a busy medical setting you don't have a lot of time and I think it's taking that into consideration as well, particularly if you are on a round or about to see someone and you have who knows how many patients to get through, then you have to be able to scan through a set of notes, pick out the key points in the history, think about that, interpret it*

and then go and make your decision and further assessment from that. So I think it's all well and good to sit for half an hour reading something but actually, you don't realistically have the time.

FP6 *In terms of practical application as well, there is a very big difference between sitting down and doing the test in a formal environment than to actually having to read something on the fly. Having a busy day in an emergency room is totally different, it's not like they can sit down in a nice quiet area and think about it for 20 minutes.*

FP7 *I think if we were to recommend that it was raised, I suppose you would hope that if you have a higher level on the test then that is going to…, like when you are in a more demanding, stressful situation, that somebody who is perhaps just scaring in on a 7 or 7.5 who may do OK on the test but actually when they are put into that more demanding environment may struggle more than someone who comes in with a higher basic level. So I would think it makes sense to raise it.*

FP8 *The reading side of things will in many ways be more technical reading reports and things but there will also be written communication from the patients and it's important that IMG can read it, comprehend it, pick out the nuances and my inclination would be that going up to 8 is probably not a bad thing because from what I remember of the test it was very much a kind of comprehension test.*

FP2 *So in essence you are saying that just getting a standard reading test which is not medically specific, the higher the skills the better the chances will be when it comes to whatever you are reading. So we are saying that we want really good skills for reading.*

Following discussion of the reading test, it was agreed that the required band score for reading should be Band 8.

Comparisons were then offered regarding the relative merits of the **reading and** the **listening** tests:

FP8 *I think the reading test is probably a better test than the listening test because I don't think the listening test really picks up on any of the 'can do's.*

Mod *I have all your comment from the earlier panels. Everybody seems to think that the reading was much more difficult than the listening, that it's just on a different level altogether.*

A number of comments were then made relating specifically to the format and content of the **listening** test, followed by comments relating to what is actually needed in terms of listening, culminating in factors such as ensuring that patients have confidence in their doctors:

FP7 *It was really measured and quite slow and structured and bland accents.*

FP8 *Well the listening was just a joke to my mind. It was just sitting down listening to what was said and writing down what was said in a box. It didn't show any need to comprehend or understand what has been said and interpret that, which as a doctor is what you have to do. You have to listen to what someone is saying to you, pick up on any undercurrents and hidden agendas and all the stuff they teach*

you in GP and then be able to use the information that you've got. It's not just a case of someone saying "I have this, this, this" and writing down "that, that, that and that".

FP6 It isn't really just about comprehension… and you know when you're seeing a patient, it's so much more than someone listing a set of symptoms, a lot of subtlety and meaning can be conveyed in the tone of voice but from what I remember, the listening test didn't really address that.

FP7 And I think listening to someone effectively, it's not just hearing it, it's listening to it and then responding to it as well and that in itself is quite a complex skill, to be able to reflect to someone that you have actually heard them correctly and that you are clear on what they are telling you.

FP2 It's not just about saying back what you heard.

FP8 The test itself doesn't actually test any kind of active listening and interpretation in that way and I think from that point of view we should set a high minimum.

FP4 One of the things I think we agreed was that the listening skills need to be higher because listening is something you probably only get a chance to do once. The others you are more likely to be able to repeat other than in an emergency situation. But listening, people have a tendency to want to be listened to the first time.

FP7 You can ask perhaps once to have someone say something again but actually more than that it gets irritating and you lose the flow of the conversation.

FP6 And in terms of patients having confidence in their doctor, I think if you told them something and then they said "sorry, can you say that again…"

FP3 I think also about confidence in your doctor, but a lot of patients wouldn't have the confidence to say "did you understand what I am saying?" or "have you heard what I am saying?" because the power position is so different. People just feel intimidated quite often so they will come out saying "you didn't understand a word I said" but in the situation not feel able or confident to challenge it.

FP7 In terms of safety as well I would want to know that if somebody had been given an instruction to do something, that they understood it correctly and were then going to go off and do it properly and hadn't sort of thought I heard something, jot it down, and then gone off and carried out what they thought were the orders. That is frightening.

FP4 That might be why we all scored it higher because we felt maybe that it wasn't discriminatory enough.

FP8 And that's why we put it so high; we didn't think it would discriminate to any real degree unless you had it at the highest level.

FP7 I think it's clear from all of the listening that 8.5, that is the only one that is a definitive agreement, isn't it, between all of the groups that listening is the key and I do agree with that. Most of communication is about listening and taking in the information and particularly in that patient setting.

Following discussion of the listening test, it was agreed that the required band score for listening should be Band 8.5.

Productive skills – Writing and Speaking

It will be remembered that the most appropriate method for determining which band scores to recommend for the productive skills was an examinee-centred approach in which judges make decisions based on test task performance (recordings of speech or copies of written work). Despite the fact that the only writing script judged to be clearly acceptable was Script 7 (standardised as representing an ability level of Band 8.5) there was very little discussion as to which Band score to recommend to the GMC for writing. Participants on the final panel had originally been on panels where decisions on a recommended writing band had ranged from Band 6 to Band 9+ (which, of course, does not exist), giving further confirmation of the difficulty of achieving rater agreement, identified by Eckes (2009) in a general writing context and Banerjee (2004) in this specific writing context, and previously mentioned in Section 4.5. As pointed out in 4.4.3., participants found Script 3 (representing Band score level 7.5) the most difficult script to judge and it was ultimately rejected as unacceptable by all categories. Nevertheless, as every initial category of participant had recommended Band 7.5 for writing, this was accepted by the Final Panel, without further discussion, as the Band score that should be recommended for Writing:

FP7 The writing is a definitive agreement, isn't it, between all the groups.

A lengthy discussion then ensued regarding the minimum band score level that should be recommended for speaking, starting with a comment by a member who had been on a doctors initial panel seeking to understand why doctors had awarded a lower Band score requirement than other categories:

FP8 We scored the speaking much lower than everyone else had. We said 7 was acceptable. In my area we recruit a lot of IMGs and we wondered actually if our experience with working with IMGs is affecting our perception of what you need to be able to do for the test. We are hearing somebody at Band 7 on this test and we are saying "we know people like that" and maybe that was affecting our scoring of it.

FP6 Could that be the other side of the coin, the fact that you have worked with IMGs and you know that practicing all these skills in a medical environment will improve it? Which is great but I think as a patient you still want a minimum level of competency.

FP4 If you meet that doctor in the first week of the four years as a patient, the fact that they are going to be better in 3 years' time is a problem, isn't it?

FP3 A lot of people's experience is with a junior doctor. That is who you see. So I think that the level should be at a higher level.

FP5 It was felt on balance by all the panels that although they judged Band 7.5 to be adequate, of the actual candidates they listened to, they only deemed the ones who got Band 8 to be successful. So, in other words, there is a conflict between the ways the actual marker was felt to be acceptable versus the candidates that were

deemed to be successful. Personally I would say there should be a minimum score of Band 8.

FP4 *Well, Band 8 was universally deemed as being adequate, the 7.5 got a mixed response.*

FP2 *Is your point going to be that, that a proviso has to go in there to say that it's not evidenced by the mean score, however this is an anomaly of it that came out for your attention?*

FP5 *Yes, the mean scores indicate 7.5 to be adequate but actually if you are looking down at how the information was derived, the anomaly is that everyone felt that Band 8 was deemed to be the minimum in terms of the actual candidates. But for some reason, the average score worked out at 7.5.*

A question was then directed to the moderator in an attempt to understand the anomaly of the judgment versus the mean score.

FP5 *Doctors seem to have lower scores for speaking. Was there a particular outlier who made that score so low and artificially lowered the mean?*

Mod *That is exactly what happened. We had someone very strong-minded, very senior, who absolutely insisted on recording 6.5 which, as you say, artificially lowered the overall mean.*

FP5 *Which is why the score for doctors is 7.*

The discussion again returned to the patients' (and nurses') views and was considered to provide a very good justification for recommending Band 8 for Speaking.

FP4 *So that's why it might have come out at 7.5, but the justification, if you like, for changing it to a Band 8 is probably best put that actually we believed the patients' view was more important than the professionals' view and therefore because the patients wanted Band 8 for speaking, that's why we said it. Justifying it by saying that actually the patients would feel more comfortable.*

FP7 *You could say lending more weight to patients. That group scored 8.*

FP5 *The nurses and patients, those who have more to do with listening to the doctors might carry a greater weight.*

FP4 *That is a very good justification.*

Interestingly, without using the actual words, the final justification offered for specifying Band 8 as the required speaking score was in complete accord with the method of standard setting identified as appropriate for dealing with the productive skills:

FP4 *The evidence you have for putting 8 for speaking is the actual conversation you have from everyone involved and the breakdown of the scoring that the panels were only happy with the four candidates that scored 8 and above. And that is very clear evidence.*

Overall Band score requirement

Discussion continued regarding the overall band score requirement which proved difficult and unpopular to agree on as the majority opinion was that a profile of different scores would be the best option. The scores required were summarised by the moderator as follows:

Mod *You don't want a clean, straightforward score, you want a profile. You want, Band 8 overall and for speaking and reading and 8.5 for listening and nothing must be below 7.5.*

The vagueness with which the reading score level had originally been specified as Band 8 when discussing the receptive skills was now forgotten, despite the fact that both the nurses' and patients' had asked for Band 8 and the following comment had not been contradicted:

FP8 *…. my inclination would be that going up to 8 is probably not a bad thing ….*

It would seem that the discussion of the receptive skills had been much more focused on listening than reading and in the productive skills, much more on speaking than writing as this comment contradicting the moderators' summary illustrates:

FP8 *I think if you would say 8 for speaking, 8 overall and 8.5 for listening with nothing below 7.5 that would cover what is here.*

FP4 *But it allows a degree of flexibility.*

FP7 *And also as we said earlier with the reading and writing those are both skills that you can do slightly at your own pace, whereas the listening and speaking are the more immediate skills that actually you hope they have better fluidity in.*

The final decision on scores required was re-summarised by the moderator as follows:

Mod *OK, just let me make sure I have this right. Overall Band 8 and speaking must be Band 8. Listening must be Band 8.5 and nothing must be below 7.5.*

FP4 *Yes, and that's exactly what the patient groups[10] asked for if you look. And I still think that is a very powerful justification for the conclusion.*

FP5 *I think because it's identical for the nurses who deal with foreign doctors all the time, that is an added justification.*

FP4 *So basically 35 out of the 62 participants asked for these scores. These scores reflect the wishes of more than 50% of the participants. So the more you look at it the more justifiable it is.*

e) The social responsibility of raising IELTS requirements

The panel deliberated the consequences of raising the IELTS band score requirements and the social consequences for people who could not meet the higher demands. After much discussion, the consensus view is encapsulated in the following comment about taking the IELTS test overseas before coming to the UK.

[10] This is not strictly true as the nurses' and patients' groups asked for a profile of: Band 8 overall, reading and speaking; Band 8.5 for listening and Band 7.5 for writing.

FP8 Thinking about the people taking the test in their home country wherever you are from and then people would come over here and be disappointed when they don't pass the PLAB and they are told they can't practise in this country. And actually it's better for them to have a harder test. I think you need to avoid that disappointment but also you need a screening test to say, oh well it's not going to be suitable for you to come, but something that people can take at home to see whether they can then come and take the thing to avoid that kind of unnecessary travel.

f) IELTS as a test of language proficiency for doctors.

The discussion addressed the perceived inadequacy of the listening test, which featured in all initial panel discussions and continued in similar vein in the final panel. However it was not thought practical or reasonable to suggest that the listening test should be replaced, as the following comment illustrates:

FP8 I think to say that any part of the test is unacceptable, we are looking at the test from a slightly different perspective to what the test was designed for, as a test of language and language skills. When everyone went through the test they were thinking if it was acceptable for a doctor, the test isn't particularly acceptable. As a test of language, it is. It's just that kind of next step on and if we're saying this is just a screening test then it's an acceptable test but we need to expect the highest mark.

This was followed by a consideration of the IELTS test as a whole and its suitability to assess the language competence of doctors. As several members of the panel were familiar with the PLAB test, the two tests were inevitably linked together, despite several reminders that the PLAB test was outside the remit of this study. There seemed to be some confusion as to whether the IELTS test assesses communication skills, which of course it does, but in the general sense of testing communicative language ability, rather than testing medical communication skills, which is the responsibility of PLAB. The general consensus seemed to be that the IELTS test provides a perfectly good assessment of general language ability but that as doctors need to communicate in medical settings, IELTS ought to be supplemented or followed by a more specific, medically-oriented test:

FP8 Something like IELTS as a screening test is good but it's the first step in a process. It shouldn't be the only step.

FP5 As a screening test I think it does work but again it's a test to see if you're ready for the next test, in and of itself maybe it's not that useful but it's certainly a screening.

FP8 We said that IELTS as a test of medical communication isn't really up to scratch but as a general test it's fine.

FP4 If they accept what we are saying about IELTS then the next stage is to say, well, what is the correlation between IELTS scores and current PLAB success? If by re-setting the IELTS required fewer doctors would spend the money to come here and fail PLAB, basically that is quite a good justification for not changing PLAB. But if they find that PLAB and IELTS don't really match then actually they have to change PLAB anyway because that demonstrates that it isn't fit for purpose.

Because we are sort of accepting that IELTS is fit for purpose as a basic screening tool for English language skills.

The panel were asked to consider if the IELTS test by and of itself provides an adequate measure of English language ability for overseas medical practitioners seeking admission to the GMC register. The response was unanimous:

All *No.*

It is clear from the chorus of the entire panel in the above response that some confusion still existed as to the separate purposes of the IELTS test and the PLAB test. It is possible that the phrasing of the question confused the panel who perhaps thought that it was being suggested that IELTS could become the only test required for registration:

FP4 *Do you think they have asked that question, or that question is part of your research, because they were assuming that we would ask for a higher IELTS band and then they could get rid of PLAB altogether? Because it that's the reason for asking the question then clearly the answer is no. We have to have another method of assessing language skills. IELTS is a screening test, we have agreed on that and suggested bands to be attained as part of the screening test, but that is all we can see IELTS can be used for.*

However, it is obvious that although there were serious concerns from the final panel and from most of the initial stakeholder panels about the suitability of the listening component of the IELTS test, as reflected in the comments above and in Appendix 2 (which will not be repeated here), the main concern amongst the final panel members seemed to be that the IELTS test doesn't test the key communication skills that a doctor needs to display in a medical setting. This is not what the IELTS test was originally designed to do and is not, of course, what it claims to do; the purpose of the IELTS test is to assess a candidate's ability to use language for the purpose of achieving a particular communicative goal in a specific situational context in each of the four skills of reading, writing, listening and speaking, discretely. Conversely, the specific purpose of the PLAB test is to assess a candidate's ability to use appropriate communication skills in specific medical and clinical contexts. This confusion may have been responsible for some of the following comments:

FP6 *As a screening to progress to the next level of assessment regarding competence, yes, it probably is. But in terms of the test in and of itself and the skills that it is assessing, no.*

FP8 *In and of itself it is not adequate, but as a first step, yes, but there needs to be other tests of language, be it expanded PLAB, be it whatever.*

FP5 *I personally think we ought to be looking at extending PLAB if it's possible to different levels of skill. I don't know how that would be achievable.*

Suggestions were then made as to what sort of skills tests would be appropriate for doctors. There was general agreement that there should be emphasis on the integrated nature of the skills that doctors need to use, something that was very strongly suggested by the initial stakeholder panels.

FP7 It would be interesting to make it as you say a combined thing so that you have perhaps testing in all areas but then you have one section that combined everything. They have to read something, write something down at the same time they are listening and having a conversation to put it all together a bit more.

g) Requiring evidence of English language ability from both IMG and EEA doctors.

There was general agreement that, regardless of EU law, all doctors intending to practise in the UK should provide evidence of English language ability before being allowed to register with the GMC. There was also discussion of the new regulations that will require every doctor to undergo revalidation by an RO and this was proposed as a possible way of assessing the language skills of all doctors, UK graduates included.

FP6 Would an RO be more conscious of the linguistic skills of a foreign doctor than they would for a native?

FP4 I think they will apply the same standard to everybody; there is no potential for being specific. You don't necessarily know just by looking at somebody or by reading somebody's name where they qualified from anyway. There are an awful lot of British graduates who don't have traditionally Anglo-Saxon names.

FP8 The GMC is responsible for fitness to practise and licence to practise and if you can make it a requirement of registration that there is a demonstration of competence in English and you can probably specify GCSE grades or IELTS score for people who don't have that GCSE, by making it a fitness to practise issue you can get round EU law that way but only if it's applied to everyone.

FP4 That's why the RO and medical directors' panel would have no problem applying it to everyone.

FP5 The bottom line is that IMGs will be asked to take the IELTS test with the previously specified banding levels requirements and then go on to take the PLAB and then to enter revalidation. The EEA doctors and those who were previously IMGs but became EEA doctors because of EC rights, we are asking that the GMC will take on a function to make it a mandatory requirement that somehow they provide alternative proof to IELTS at Band 8.

FP8 Ultimately this is all about patients' safety and it's a step in the right direction if any doctor who doesn't have English as their first language, if they have to take an English language test, that is at least a step in the right direction in trying to make things safer for patients. And if by introducing IELTS for EEA doctors screens out a handful who otherwise would have had poor communication and made mistakes as a result of that, and it means a handful of fewer problems to patients, that is a good thing. If that's what comes out of this panel then that is fantastic. And even if all we do is recommend IELTS for EEA doctors as a demonstration of English language, it's a small step but at least it's a step in the right direction.

5.2. Discussion and recommendations

This study has been undertaken in order to refresh and extend a previous investigation into IELTS Band score levels and IMGs (Banerjee 2004) and to develop a stronger

evidence base for the GMC's requirements when setting minimum language standards for overseas applicants to their register. Current required IELTS levels have been investigated to determine if they are adequate in light of issues of patient safety; the issue of requiring evidence of English language ability from all non-native English speaking medical practitioners seeking admission to the GMC register has also been investigated. The following questions emerged from the literature review on medical communication:

1. Does the IELTS offer an appropriate measure of the language ability of prospective medical professionals?
2. If it is found that IELTS is an appropriate instrument, what is the most appropriate level for prospective medical doctors to attain before being accepted to practise in the UK?
3. Should all non-native-English speaking doctors be asked to empirically demonstrate their language level?

These questions will be addressed individually and recommendations made based on the final, confirmatory panel's deliberations presented in 5.1., above.

5.2.1. Does the IELTS offer an appropriate measure of the language ability of prospective medical professionals?

As shown in the literature review (Section 2.3), assessing language ability in professional contexts is a highly complex and somewhat daunting task. In a medical context, with the exception of the Australian Occupational English Test (OET), itself a controversial and specialized performance test that is expensive to operate (McNamara, 1996:99), no other medically-oriented language proficiency test of English for doctors exists. However, a recent empirical study by Elder et al. (2012) has highlighted the critical importance of doctors having a good command of lay language, in addition to medical terminology. Wette (2011 and Pill and Woodward-Kron's (2012) response to Wette, have also highlighted the tensions that exist between the assessment of general language proficiency and medical communication skills It appears, then, that in addition to having acceptable medical qualifications, prospective doctors should be able to demonstrate a high level of general English language proficiency.

Despite the criticisms made of the IELTS test, which were mainly concerned with the appropriacy of using a test designed for use in the academic domain in other domains (McNamara, 2000; McNamara and Roever, 2006), IELTS has been shown to be a robust general language proficiency test. This is reinforced by comments from the initial stakeholders panels and from the Final Panel (see Appendix 2).), most of whom have criticised certain aspects of the test but most of whom also consider it to be a very good test of general English ability. It is clear, therefore, that although IELTS obviously cannot be used as a test of medical competence, it is certainly an appropriate instrument to use to assess the language ability of prospective medical doctors to the GMC register.

RECOMMENDATION 1: In light of the demonstrated need for medical professionals to have a high level of general English language ability in addition to accepted medical qualifications, we recommend that the IELTS test should be retained as an appropriate test of the English language competence of overseas-trained doctors.

5.2.2. What is the most appropriate level for prospective medical doctors to attain before being accepted to practise in the UK?

As stated earlier, there are a number of alternative pathways available for fulfilling GMC requirements for registration to practice medicine in the UK. However, the current most popular way is for IMGs to provide initial evidence of English language competence by achieving a minimum overall score of Band 7 on the academic module of the IELTS test in a single sitting, with no separate skill score lower than Band 7. On achieving this, they are then eligible to register to take Part 1 of the Professional and Linguistics Assessment Board (PLAB) test which they must pass in order to be eligible to take Part 2 of the PLAB, success in which makes them eligible for registration with the GMC which, in turn, allows them to practice medicine in the UK.

In view of the concerns expressed in 5.2.1 above, with regard to only using IELTS as a screening device for language proficiency and having a separate test of medical communication skills, there is obviously no band score level that could possibly be obtained on the IELTS test that would automatically enable a prospective medical doctor to be accepted to practise in Britain. However, the question can be more adequately addressed if rephrased as follows:

> *Is the current overall IELTS score of Band 7, with no separate skill score lower than Band 7, adequate as a preliminary language screening device for IMGs?*

As can be seen from the comments presented in 5.1.3.d., Band 7 is not considered to represent an adequate overall band score, nor even an adequate band score for any individual skill. The Final Panel, in keeping with several of the initial stakeholder panels, found it very difficult to arrive at an overall band score to recommend, many participants preferring to specify a profile. However, as an overall band score is an essential component of IELTS reporting, it was eventually agreed that an average of the profile scores would be recommended (which is how the overall IELTS score is computed) which would give an overall Band score requirement of Band 8.

As mentioned previously, the final panel did not want to recommend a flat rate of Band 8 (such as exists at present with Band 7). This is supported by research findings such as the study by Chalhoub-Deville and Turner (2000) who reviewed all the major tests used around the world for undergraduate and postgraduate admissions to university. Chalhoub-Deville and Turner (2000: 537) suggest that good selection practice should take into account not only an overall band score, but also scores in the different skills areas, as different academic programs may require different profiles of language ability. Although this study is not concerned with university admissions, it could be argued that there may be more demand on doctors' oral/aural skills than on their receptive and productive written ability. Consequently, the GMC may wish to take this into consideration when determining the English language requirements for overseas applicants to their register.

The final panel considered a number of pathways that could potentially be adopted to achieve an overall band score of Band 8 but ultimately the choice was only between the following three:

1. Require the scores that were originally agreed by the Final Panel with no variation in profile allowed, namely Overall - Band 8, Listening Band - 8.5,

Speaking - Band 8, Reading - Band 8, Writing - Band 7.5. This provides an exact overall average of Band 8.

2. Accept, as is commonly stated and supported by research (McNamara 1996, Elder et al 2012, *inter alia*), that oral skills are the most important but allow some flexibility on the reading and writing scores so that ONE of the two should be Band 8 but the other could be Band 7.5 (without specifying which). This would also yield an overall band score of Band 8, but was eventually rejected as being far too complicated to explain.
3. Accept that oral skills are the most important as in 2 above, but allow flexibility on the reading and writing scores, as subsequently suggested by the Final Panel, and require Overall - Band 8, Listening Band - 8.5, Speaking - Band 8, but allow both Reading and Writing to be Band 7.5. This would yield an overall band score of 7.875, which would, of course, be rounded up to Band 8.

RECOMMENDATION 2: We recommend that the band score level requirements for IELTS should be revised and that the GMC should consider adopting the following profile which reflects the importance of oral skills, with listening being of paramount importance, but allows for some flexibility in assessing written skills:

Overall	Band 8
Listening	Band 8.5
Speaking	Band 8
Reading	Band 7.5
Writing	Band 7.5

5.2.3. Should all non-native-English speaking doctors be asked to empirically demonstrate their language ability?

Despite an EU directive which currently exempts graduates with acceptable medical qualifications from within the European Economic Area and Switzerland from providing evidence of English language ability when applying for registration in the UK, this is not considered an acceptable situation with regard to patient safety. As can be seen from the comments presented in 5.2.1.g., it is forcefully argued that both IMGs and EEA graduates should provide evidence of English language ability in order to register with the GMC.

RECOMMENDATION 3: We recommend that the GMC should attempt to find a way of requiring all non-native speakers of English to provide evidence of English language competence before being allowed to practise medicine in the UK and that if ever it becomes possible to require evidence of language ability from EEA graduates, they should provide the same evidence as IMGs.

5.3. Limitations of the study, further discussion and suggestions for consideration

5.3.1. Limitations of the study

Ideally, for the study to be considered truly representative of the population of the UK as a whole, it would have sampled initial stakeholders (i.e. patients, nurses and doctors) from every region of the UK, i.e. from each of the north, south, east and west of England,

plus Northern Ireland, Scotland and Wales. Even if we had only increased the number of AHP and RO panels to include one each from England, Northern Ireland, Scotland and Wales, this would have meant conducting a total of 29 initial stakeholder panels. This would have been problematic for a number of reasons:

1. It would have vastly increased the cost and therefore the budget allocated for this research.
2. It would have made the workload impossible for the existing team of researchers, which would have meant recruiting additional members. One of the advantages of the study as it stands is that the same researchers participated in every panel and could thus draw conclusions based on experience. A diverse group of researchers would not have had that benefit.
3. Given the problems encountered in trying to recruit members to the 11 panels we conducted, recruiting participants in 29 panels in 7 different locations, may have posed a considerable logistical challenge.
4. The length of time spent transcribing the discussions and analysing the qualitative data would have increased enormously, thus postponing submission of the final report by some considerable time.

In terms of the materials used, we were very fortunate to have broad cooperation from Cambridge ESOL who provided all the materials we requested. However, all the writing samples and speech exemplars were chosen by Cambridge ESOL; it might have been useful to be able to select our own exemplars from a wider range. It would also have made discussing the writing and to some extent the speaking samples easier if we had been given access to the actual marking band sheets used for scoring rather than the publicly published ones. In terms of the speaking exemplars, it would have been useful to have also been able to include samples of real, overseas-trained doctors.

This research was conducted in isolation from any organised medical contexts. Given the well-documented, highly interactive nature of communication in medical settings and the fact that there already exists a test of medical communication skills for overseas –trained doctors as well as numerous specialist communication skills test such as that for GPs, it would have perhaps been useful to cooperate with medical professionals to see if we could contribute to each other's understanding of necessary language competence and communication in a medical context.

5.3.2. Further discussion and suggestions for consideration

In terms of IMG candidates applying for admission to the GMC register, it may be that language is not the only issue and perhaps there should be a closer look at the relationship between language ability as assessed by IELTS and medical knowledge as assessed by PLAB. Bearing this in mind the comment reprinted below offers a useful suggestion:

FP4 If they accept what we are saying about IELTS then the next stage is to say, well, what is the correlation between IELTS scores and current PLAB success? If by re-setting the IELTS required fewer doctors would spend the money to come here and fail PLAB, basically that is quite a good justification for not changing PLAB. But if they find that PLAB and IELTS don't really match then actually they have to change PLAB anyway because that demonstrates that it isn't fit for purpose.

Because we are sort of accepting that IELTS is fit for purpose as a basic screening tool for English language skills.

SUGGESTION 1: In order to determine whether or not there is a strong relationship between performance on the IELTS test and success in each part of the PLAB test, we suggest that the GMC consider statistically analysing the relationship between score performance on IELTS and score performance on PLAB Parts 1 and 2.

It may also be useful to consider the argument that the IELTS test, particularly the listening test, does not provide adequate evidence of the skills previously identified as being essential for a doctor (as outlined in 4.5 for Listening and in the 'can do' statements in Appendix 1 for the entire test) .

Several professions where language skills are considered critical have considered this and have commissioned specific tests to meet their individual requirements. The test for 999 telephone operators was mentioned in 4.1; a further example of a profession-specific test is the test developed for air traffic controllers (see http://www.eurocontrol.int/services/first-european-air-traffic-controller-selection-test-feast).

Discussing the Canadian medical context, Watt et al. (2003:35) recommend that Canada works to develop and implement "a national language standard and standard assessment procedures for demonstrating the language proficiency required for medical practice in Canada" and suggest that the language standard be one that "parallels the existing Canadian Language Benchmarks (CLB), and is adjusted to more accurately reflect the language demands of working in medical contexts".

There is absolutely no reason why the same could not also be done for the UK as the UK Occupational Language Standards already exist and could readily be adapted and extended to cater for medical practice (see http://www.cilt.org.uk/home/standards_and_qualifications/uk_occupational_standards/languages.aspx). In addition, this research has identified numerous skills that are considered essential for medical practitioners (see Appendix 1). These skills are linguistic in nature and are not necessarily comparable to the medical communication skills taught on undergraduate and graduate medical courses in the UK. Arguing for greater cooperation between medical professionals and applied linguists in the assessment of doctors and health professionals, Elder et al. (2012:417) comment:

It is certainly the case that models of communicative competence informing practice in language testing are different from the views of communication informing the HP communication literature, and hence the views of educators in that field. Rapprochement between these two perspectives is clearly desirable. Just as applied linguistic views of communication would become richer and more relevant, so the teaching of clinical communication skills might be more usefully informed by greater attention to language.

SUGGESTION 2: We suggest that the GMC considers developing a UK language standard and standard assessment procedures for demonstrating the language proficiency required for medical practice in the UK.

5.4. Conclusion

A further recommendation from Watt et al. (2003), that the Canadian Language Benchmarks Assessment (CLBA) be added to the list of tests, while retaining international tests such as TOEFL and IELTS, is an indication of their awareness of the importance of the local context in test validity, as it implies that the experiences of the candidates within the context of the medical domain are likely to play an important role in our understanding of their language needs within that domain. By investigating the use of IELTS as a screening tool for overseas applicants to the GMC register, we believe that the research presented in this report goes some way towards building a substantive validity argument for the use of IELTS in the context of the UK medical domain.

6. References

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

Ahlstrum, S. & Bernspang, B. (2003). Occupational Performance of Persons Who Have Suffered a Stroke: a Follow-up Study. *Scandinavian Journal of Occupational Therapy*, 10: 88 - 94.

Ali, N. (2003). Fluency in the consulting room. *British Journal of General Practice* 53(492), 514–15.

Atchison, B. T., Fisher, A. G. & Bryze, K. (1998). Rater reliability and internal scale and person response validity of the School Assessment of Motor and Process Skills. *American Journal of Occupational Therapy*, 52: 843-850.

Baker, D. and Robson, J. (2012). Communication training for international graduates. *The Clinical Teacher*, 9, 325-329.

Banerjee, J. (2004). *Study of the minimum English language writing and speaking abilities needed by overseas trained doctors*. Report to the General Medical Council. July 2004.

Banerjee, J. and Taylor, L. (2005). Setting the standard: what English language abilities do overseas trained doctors need? Paper presented at the *Language Testing Research Colloquium*, Ottawa, Canada, July 2005.

Beck, C. T. & Gable, R. K. (2000) Postpartum Depression Screening Scale: Development and Psychometric Testing. *Nursing Research*, 49: 272-282.

Beck, C. T. & Gable, R. K. (2001a) Further Validation of the Postpartum Depression Screening Scale. *Nursing Research*. 50:155-164.

Beck, C. T. & Gable, R. K. (2001b) Comparative Analysis of the Performance of the Postpartum Depression Screening Scale With Two Other Depression Instruments. *Nursing Research*. 50:242-250.

Beck, C. T. & Gable, R. K. (2003) Postpartum Depression Screening Scale: Spanish Version. *Nursing Research*. 52:296-306.

Bernspang B. (1999). Rater Calibration Stability for the Assessment of Motor and Process Skills. *Scandinavian Journal of Occupational Therapy*, 6: 101-109.

Bernspang, B. & Fisher, A. G. (1995a). Validation of the Assessment of Motor and Process Skills for use in Sweden. *Scandinavian Journal of Occupational Therapy*, 2: 3-9.

Bernspang, B. & Fisher, A. G. (1995b). Differences between persons with right or left CVA on the Assessment of Motor and Process Skills. *Archives of Physical Medicine and Rehabilitation*, 76: 1144-1151.

Berry, V. (2007). *Personality Differences and Oral Test Performance*. Frankfurt am Main. Peter Lang.

Biddle, R. (1993). How to Set Cutoff Scores for Knowledge Tests Used In Promotion, Training, Certification, and Licensing. *Public Personnel Management*, 22(1): 63-70.

Bode, R. K., Klein-Gitelman, M. S., Miller, M. L., Lechman, T. S. & Pachman L. M. (2003). Disease activity score for children with juvenile dermatomyositis: Reliability and validity evidence . *Arthritis Care & Research*, 49: 7-15.

Bonk, W.J. and Ockey, G.J. (2003). A many-facet Rasch analysis of the second language grouop oral discussion task. *Language Testing*, 20 (1): 89-110.

Brown, A. (2005). *Interviewer Variablity in Oral Proficiency Interviews*. Franfurt am Main. Peter Lang.

Campbell, S. K., Kolobe, T. H. A., Wright, B. D., Linacre, J. M. (2002) Validity of the Test of Infant Motor Performance for prediction of 6-, 9- and 12-month scores on the Alberta Infant Motor Scale. *Developmental Medicine & Child Neurology*, 44: 263 – 272.

Candlin, C.N. and Candlin, S. (2003). Healthcare communication: a problematic site for applied linguistic research. *Annual Review of Applied Linguistics* 23, 134-154.

Cegala, D. J., Coleman, M. T. and Turner, J. W. (1998). The development and partial assessment of the medical communication competence scale. *Health Communication* 10(3), 261 – 288.

Chalhoub-Deville, M. and Turner, C. (2000). What to look for in ESL admissions tests: Cambridge certificate exams. IELTS and TOEFL. *System*, 28: 523-539.

Chang, L. (1999). Judgmental Item Analysis of the Nedelsky and Angoff Standard-Setting methods. *Applied Measurement in Education*, 12(2): 151-165.

Chien, C-W., Brown, T. & McDonald, R. (2012). Examining construct validity of a new naturalistic observational assessment of hand skills for preschool- and school-age children. *Australian Occupational Therapy Journal*, 59: 108–120.

Chur-Hansen, A. (1997). Language background, English language proficiency and selection for language development. *Medical Education*, *31*, 312-319.

Cizek, G. J. (2011). *Reconceptualizing validity and the place of consequences.* Paper Presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, April 2011.

Cizek, G. J. and Bunch, M. B. (2007). *Standard Setting*. Thousand Oaks, CA: Sage.

Cizek, G.J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement,* 30(2), 93-106.

Cohen, A.L, Rivara, F, and Marcuse, E, K. (2005). Are language barriers associated with serious medical events in hospitalized pediatric patients? *Pediatrics* 116(3), 575–579.

Darragh, A. R., Sample, P. L. & Fisher, A. G. (1998). Environment effect of functional task performance in adults with acquired brain injuries: use of the assessment of

motor and process skills. *Archives of Physical Medicine and Rehabilitation*, 79: 418-23.

Davidson, B. (2000). The interpreter as institutional gatekeeper: The social-linguistic role of interpreters in Spanish-English medical discourse. *Journal of Sociolinguistics* 4(3), 379 - 405.

de Morton, N. A., Keating, J. L. & Davidson, M. (2008). Rasch analysis of the barthel index in the assessment of hospitalized older patients after admission for an acute medical condition. *Archives of Physical Medicine and Rehabilitation*, 89: 641-647.

Decruynaere, C., Thonnard, J-L. & Plaghki, L. (2007). Measure of experimental pain using Rasch analysis. *European Journal of Pain*, 11: 469-474.

Divi, C. Koss, R. Schmaltz, S. and Loeb, J. (2007). Language proficiency and adverse events in US hospitals: a pilot study. *International Journal for Quality in Health Care* 19(2), 60–67.

Drew, P., Chatwin, J. and Collins, S. (2001). Conversation analysis: a method for research into interactions between patients and health-care professionals. *Health Expectations* 4(1), 58 – 70.

Eckes, T. (2009). Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment(Section H).* Strasbourg, France: Council of Europe/Language Policy Division.

Elder, C., Barkhuizen, G., Knoch,U., and von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing*, 24, 37–64.

Elder, C., Knoch, U., Barkhuizen, G., and von Randow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly*, 2, 175–196.

Elder, C., Pill, J., Woodward-Kron, R., McNamara, T., Manias, E., McColl, G. and Webb, G. (2012). Health professionals' views of communication: Implications for assessing performance on a health-specific English language test. *TESOL Quarterly*, *46*(2), 409-419.

Fitzpatrick, R., Norquista, J. M., Dawsonb, J. & Jenkinsonc, C. (2003). Rasch scoring of outcomes of total hip replacement. *Journal of Clinical Epidemiology*, 56: 68-74

Faggen, J. (1994). *Setting standards for constructed response tests*: *An overview* (ETS RM- 94- 119). Princeton, NJ: ETS.

Flores, G, Rabke-Verani J. and Pine, W. (2002). The importance of cultural and linguistic issues in the emergency care of children. *Pediatric Emergency Care* 18(4), 271 – 284.

Flores, G., Barton Laws, M., Mayo, S.J., Zuckerman, B., Abreu, M., Medina, L. and Hardt, E.J. (2003). Errors in Medical Interpretation and Their Potential Consequences in Pediatric Encounters . *Pediatrics* 111(1), 6 - 14.

Friedman, M., Sutnick, A.I., Stillman, P. L., Norcini, J. J., Anderson, S. M., Williams, R. G., Henning, G. and Reeves, M. J. (1991). The use of standardized patients to evaluate the spoken-English proficiency of foreign medical graduates. *Academic Medicine, 66,* S61–S63.

Girard, C. R., Fisher, A. G., Short, M. A. & Duran L. (1999) Occupational performance differences between psychiatric groups. *Scandinavian Journal of Occupational Therapy*, 6, 119 - 126

Haertel, E. H. (1999) 'Validity arguments for high-stakes testing: in search of the evidence.' *Educational Measurement: Issues and Practice* 18, 4, 5–9.

Hambleton, R. K., and Plake, B. S. (1995). Using an Extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education*, 8(1), 41-55.

Hambleton, R.K., Brennan, R.L., Brown, W., Dodd, B., Forsyth, R.A., Mehrens, W.A., Nellhaus, J., Reckase, M.D., Rindone, D. van der Linden, W.J. and Zwick, R. (2000). A response to "Setting Reasonable and Useful Performacne Standards" in the National Academy of Sciences' Grading the Nation's Report Card. *Educational Measurement: Issues and Practice*, 19(2), 5-14.

Hamp-Lyons, L. (2007). Worrying about rating. *Assessing Writing*, 12 (1), 1–9.

Harding, L., Pill, J. and Ryan, K. (2011). Assessor decision making while marking a note-taking listening test: The case of the OET. *Language Assessment Quarterly, 8*(2), 108-126.

Hayase, D., Mosenteen D., Thimmaiah, D., Zemke, S. & Atler, K., Fisher A.G. (2004). Age-related changes in activities of daily living ability . *Australian Occupational Therapy Journal*, 51: 192.

Heritage, J. and Maynard, D. (2006). Introduction: analyzing interaction between doctors and patients in primary care encounters. In Heritage, J., & Maynard, D. (Eds.), *Communication in medical care*. Cambridge, MA: Cambridge University Press, pp.1-21.

Hoekje, B (2007). Medical discourse and ESP courses for international medical graduates (IMGs). *English for Specific Purposes* 26(3) 327–343

Hurtz, G. M. and Hertz, N. (1999). How many raters should be used for establishing cutoff scores with the Angoff method? A Generalizability Theory Study. *Educational and Psychological Measurement, 59*(6) 885-897.

Ibey, R. J.; Chung, R., Benjamin, N., Littlejohn, S., Sarginson, A., Salbach, N. M., Kirkwood, G. & Wright, V. (2011). Development of a Challenge Assessment Tool for High-Functioning Children With an Acquired Brain Injury. *Pediatric Physical Therapy*, 22:268-276.

Impara, J. C. and Plake, B. S. (1998). Teachers' Ability to Estimate Item Difficulty: A Test of the Assumptions in the Angoff Standard Setting Method. *Journal of Educational Measurement,* 35(1): 69-81.

Jacoby, S. and McNamara, T. (1999). Locating Competence. *English for Specific Purposes* 18(3), 213 – 241.

Johnstone MJ. and Kanitsaki O. (2006). Culture, language, and patient safety: making the link. International *Journal of Quality Health Care* 18(5), 383–388.

Kaftandjieva, F. (2004). *Reference Supplement to the Preliminary Pilot version of the Manual for Relating Language examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment. Section B: Standard Setting.* Strasburg: Council of Europe.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, 64(3): 425-461.

Kantarcıoğlu, E. (2012). *A Case-Study of the Process of Linking an Institutional English Language Proficiency Test (COPE) for Access to University Study in the Medium Of English to the Common European Framework for Languages: Learning, Teaching and Assessment.* Unpublished PhD thesis, University of Roehampton, London.

Kantarcýoðlu, E., Thomas, C., O'Dwyer, J. and O'Sullivan, B. (2010). The cope linking project: a case study. In Waldemar Martyniuk (ed.) *Aligning Tests with the CEFR: Case studies and reflections on the use of the Council of Europe's Draft Manual.* Cambridge: Cambridge University Press.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, 26, 275–304.

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, 12, 26–43.

Kottorp A., Bernspang B., Fisher A.G. (2003) Validity of a performance assessment of activities of daily living for people with developmental disabilities. *Journal of Intellectual Disability Research*, 47: 597.

Kozaki, Y. (2004). Using GENOVA and FACETS to set multiple standards on performance assessment for certification in medical translation from Japanese into English. *Language Testing*, *21*: 1–27.

Ku, L. and Flores, G. (2005). Pay now or pay later: providing interpreter services in healthcare. *Health Affairs* 24(2), 435–44.

Lai, J. S., Velozo, C. A. & Linacre, J. M. (1997). Adjusting for Rater Severity in an Un-linked FIM national Data Base: An Application of the Many-Facets model. *Physical Medicine and Rehabilitation*, 11: 325-332.

Lear, D. W. (2005). Spanish for working medical professionals: Linguistic needs. *Foreign Language Annals*, *38*, 223–232.

Lepetit, D. and Cichocki, W. (2002). Teaching languages to future health professionals: a needs assessment study. *The Modern Language Journal*, *86*, 384-396.

Liao, P. M., Campbell, S. K. (2002) Comparison of Two Methods for Teaching Therapists to Score the Test of Infant Motor Performance. *Pediatric Physical Therapy,* 14, 4, 191-198.

Liao, P. M., Campbell, S. K. (2004) Examination of the Item Structure of the Alberta Infant Motor Scale. *Pediatric Physical Therapy*, 16: 31-38.

Linacre, J.M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.

Linacre, J.M. (1996). True-score reliability or Rasch validity? *Rasch Measurement Transactions*, 9: 455.

Livingston, S. and Zieky, M. (1982) *Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests*. Princeton, NJ: ETS.

Lumley, T. (2005). *Assessing Second Language Writing: The Rater's Perspective*. (*Vol.3., Language Testing and Evaluation series*). Frankfurt am Main: Peter Lang.

Loevinger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin*, 51, 493–504.

Lunz, Mary E., and Wright, Benjamin D. 1997. Latent Trait Models for Performance Examinations. In Jürgen Rost and Rolf Langeheine (Eds) *Applications of Latent Trait and Latent Class Models in the Social Sciences*. http://www.ipn.uni-kiel.de/aktuell/buecher/rostbuch/ltlc.htm

Malec J. F. (2004). Comparability of Mayo-Portland Adaptability Inventory ratings by staff, significant others and people with acquired brain injury. *Brain Injury,* 18: 563-575.

Maurer, T. J., Alexander,R. A., Callahan, C. M., Bailey, J. J., and Dambrot, F. H. (1991). Methodological and psychometric issues in setting cutoff scores using the Angoff method. *Personnel Psychology,* 44: 235-262.

McLachlan, J., Illing, J., Rothwell, C., Margetts, J.K., Archer, J. and Shrewsbury, D. (2012). Developing an evidence base for the Professional and Linguistics Assessments Board (PLAB) Test. Literature review submitted to GMC, October 2012.

McManus, I. C., Thompson, M. & Mollon, J. (2006). Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling. *BMC Medical Education,* 6: 42. Accessed November 4, 2012 from http://www.biomedcentral.com/1472-6920/6/42.

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

McNamara, T. *Language Testing*. (2000). Oxford: Oxford University Press.

McNamara, T. and Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing,* 29(4): 555-576.

McNamara, T. and Roever, C. (2006). *Language Testing: The Social Dimension*. Malden, MA & Oxford: Blackwell

McNeilis, K.S. (2001). Analyzing communication competence in medical consultations. *Health Communication*, 13(1), 5-18.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Nellhaus, J., Reckase, M.D., Rindone, D. van der Linden, W.J. and Zwick, R. (2000). A response to "Setting Reasonable and Useful Performance Standards" in the National Academy of Sciences' Grading the Nation's Report Card. *Educational Measurement: Issues and Practice*, 19(2), 5-14.

O'Neill, T. R., Tannenbaum, R. J. and Tiffen, J. (2005). Recommending a minimum English proficiency standard for entry-level nursing. *Journal of Nursing Measurement, 13*(2), 129-146.

O'Neil, T. R., Buckendahl, C. W., Plake, B. S. and Taylor, L. (2007). Recommending a nursing specific passing standard for the IELTS examination. *Language Assessment Quarterly*, 4(4), 295 – 317.

O'Sullivan, B. (2008). *Modelling Performance in Oral Language Testing*. Frankfurt am Main. Peter Lang.

O'Sullivan, B. (2009). *City & Guilds Communicator Level IESOL Examination (B2) CEFR Linking Project Case Study Report*. City & Guilds Research Report. Retrieved on 29/1/2009, from http://www.cityandguilds.com/documents/ind_general_learning_esol/CG_Communicator_Report_BOS.pdf

O'Sullivan, B. (2012). Assessment Issues in Languages for Specific Purposes. *Modern Language Journal*, 96 (Focus Issue): 71-88.

O'Sullivan, B., and Rignall, M. (2007). Assessing the value of bias analysis feedback to raters for the IELTS Writing Module. In L. Taylor and P. Falvey (Eds.),*IELTS collected papers: Research in speaking and writing assessment* (pp. 446–478). Cambridge, UK: Cambridge University Press.

O'Sullivan, B. and Weir, C. J. (2011) Language testing = validation. In B. O'Sullivan (Ed.) *Language testing theories and practices* (pp. 13-32). Oxford: Palgrave Macmillan.

Pallant, J. F. & Tennant, A. (2007). An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology*, March, Pt. 1: 1-18.

Papageorgiou, S. (2007). *Setting standards in Europe: The judges' contribution to relating language examinations to the Common European Framework of Reference*. Unpublished PhD dissertation. University of Lancaster.

Pill, J. and Woodward-Kron, R. (2012). How professionally relevant can language tests be? A Response to Wette (2011). *Language Assessment Quarterly*. 9:105-108.

Plake, B. and Impara, J. (2001). Ability of Panelist to Estimate Item Performance for a Target Group of Candidates: An Issue in Judgmental Standard Setting. *Educational Assessment*, 7(2): 87-97.

Rea-Dickins, P. (1987). Testing doctors' written communicative competence: An experimental technique in English for specialist purposes. *Quantitative Linguistics, 34*, 185-218.

Read, J. and Wette, R. (2009). Achieving English proficiency for professional registration: The experience of overseas-qualified health professionals in the New Zealand context. *IELTS Research Reports, 10*, 181-222.

Reed, D. J., and Cohen, A. D. (2001). Revisiting raters and ratings in oral language assessment. In C. Elder et al. (Eds.), *Experimenting with uncertainty: Essays in honour of Alan Davies* (pp. 82–96). Cambridge, UK: Cambridge University Press.

Roberts, C, Moss, B. l., Wass, V, Sarangi, S. and Jones. R. (2005) Misunderstandings: a qualitative study of primary care consultations in multilingual settings, and educational implications. *Medical Education* 39(5), 465–475.

Shapiro, M. M., Slutsky, M. H., and Watt, R. F. (1989) Minimizing Unnecessary Differences in Occupational Testing. *Valparaiso University Law Review* 23(3), 213 – 265.

Shohamy, E. (1995). Performance assessment in language testing. *Annual Review of Applied Linguistics*, 15, 188–211.

Silverman, J., Kurtz, S. and Draper, J. (2005). *Skills for communicating with patients* (2nd ed.). Oxford: Radcliffe.

Skelton, J. (2008). *Language and clinical communication: This bright Babylon*. Oxford: Radcliffe.

Skelton, J. Kai, J. and Loudon, R. (2001) Cross-cultural communication in medicine: questions for educators. *Medical Education* 35(3), 257–61.

Taylor, L. and Pill, J. (forthcoming 2014). Assessing Health Professionals. In A. J. Kunnan (Ed.), *The Companion to Language Assessment*. Malden, MA: Wiley-Blackwell.

van Moere, A. (2006). Validity evidence in a group oral test. *Language Testing*, 23 (4): 411-440.

Verlinde, E., de Laender, N., de Maesschalk, S., Deveugele, M. and Willerns, S. (2012). The social gradient in doctor-patient communication. *International Journal for Equity in Health*, 11: 12. Retrieved on 12/4/2012 from http://www.equityhealthj.com/conent/11/1/12

von Fragstein, M., Silverman, J., Cushing, A., Quilligan, S., Salisbury, H. and Wiskin, C. (2008). UK consensus statement on the content of communication curricula in undergraduate medical education. *Medical Education, 42*(11), 1100-1107.

Watt, D., Lake, D., Cabrnoch, T. and Leonard, K. (2003). *Assessing the English language proficiency of international medical graduates in their integration into Canada's*

physician supply. Report commissioned by the Canadian Task Force on Licensure of International Medical Graduates, Ottawa, Ontario, Canada. Retrieved from http://www.m-cap.ca/pdf/Assessing%20ELP%20of%20IMGs_Final_2003.pdf.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, 15, 263–287.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, 6, 145–178.

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.

Whelan, G., McKinley, D., Boulet, J., Macrae, J. and Kamholz, S. (2001). Validation of the doctor–patient communication component of the Educational Commission for Foreign Medical Graduates Clinical Skills Assessment. *Medical Education* 34(8) 757– 761.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, 10, 305–335.

Wodak, R. (2005). Critical Discourse Analysis and the Study of Doctor-Patient Interaction. In Gunnarsoon, B. L., Linell, P., and Nordberg, B. (Eds.) *The Construction of Professional Discourse*. London: Longman, 173 - 200.

Wood, A. and Head, M. (2004). Just what the doctor ordered: the application of problem-based learning to EAP. *English for Specific Purposes* 23(1), 3-17.

Wright, B, & Linacre, J. 1994. Reasonable Mean-square Fit Values. *Rasch Measurement Transactions*, 8(3): 370.

7. Appendices

Appendix 1: 'Can do' statements

Reading

Can read and understand handwritten, typed or online:

- Prescriptions
- Labels
- Signs
- Drugs charts
- Dosage regulations/protocol guidelines
- X-rays/radiography
- Laboratory reports
- Blood test results/Results of other investigations
- Emergency procedures/ advanced life support algorithms
- Abbreviations (UK, as different in all countries)

- Medical records/patients' clinical notes
- Referrals
- Letters/notes from other doctors/hospitals
- Letters/written communication from patients
- Discharge and clinic letters
- Medical legal reports
- Forms/complaint forms
- Medical journals/textbooks/dictionaries
- The BNF and BNF for Children (British National Formulary)
- Newspapers/leaflets/posters

- Trust policies/procedures/protocols – hard copy and on Intranet
- Contracts outlining their responsibilities
- Documents phrased in regulatory terms
- Hospital management minutes/directives

Can:

- skim long texts quickly for essential information
- understand lengthy, complex instructions including details on conditions and warnings, provided difficult sections can be reread.
- understand in detail a wide range of lengthy, complex texts likely to be encountered in professional life, identifying finer points of detail
- demonstrate comprehensive understanding of personal messages in informal letters, emails etc.
- when working under time pressure, demonstrate broad understanding of texts conveying detailed, new information

Writing

Can handwrite, type or dictate into a dictaphone for later transcription clear, concise, accurate and systematic:

- Patients' records, chronological histories, case notes, ward round notes
- Prescriptions
- Notes to colleagues
- Clear, precise instructions/orders to nurses
- Clear, precise instructions to patients about drug protocols
- Flow charts
- Sick notes to employers/ for benefits
- Forms, death certificates
- Appointments
- Handover of patient care
- Communicate with labs. to get results
- Communicate with patients by email
- Translate findings from patient contact into writing

- Referral letters
- Letters/notes to other doctors/hospitals
- Letters/other forms of written communication to patients
- Discharge and clinical letters
- Medical legal reports/statements to lawyers
- Police/coroners/court reports
- Detailed letters to insurance companies
- Letters to management re: red tape
- Letters to PCTs to obtain non-routine medications
- Writing therapy letters (in psychiatry)

Can:

- make full and accurate notes and continue to participate in a meeting or consultation
- express themselves with clarity and precision in personal correspondence, using language flexibly and effectively
- handle a wide range of routine and non-routine situations in which professional services are requested from colleagues or external contacts
- write a set of instructions with clarity and precision
- describe and interpret empirical data on professional topics for a general audience
- transfer information from one place to another
- interpret what they have read and rewrite it for different audience

Listening

Can listen and pick out relevant and important information from:

- Patients: adults/children/elderly/disabled/brain damaged/otherwise impaired/drunk
- Families of patients/carers
- Other doctors /senior and junior colleagues/GPs
- Nurses/Midwives
- Social workers/child protection workers
- Physiotherapists/dieticians/language therapists/paramedics/ radiographers/lab. technicians
- Police/solicitors/coroners
- Pharmacists/drug companies' representatives
- Management/Admin. staff
- Seminar/conference speakers

Can pick out relevant and important information in:

- Emergency situations while doing other things
- Excitable or stressful situations with considerable background noise
- Handovers/ward rounds/multi-disciplinary meetings/discharge situations
- Seminars/conferences
- Conversations by phone/Skype/videolink/conference calls

Can:

- follow extended, natural speech even when it is not clearly structured
- extract the gist of what is said in informal meetings and discussions involving multiple participants, marked by colloquialisms and overlapping turns
- extract the gist of what is said when conversation is animated and delivered at a fast natural rate in a range of accents.
- extract gist, detail, purpose and main points from formal discussions involving detailed information such as facts or definitions related to professional topics
- integrate information from multiple sources and follow detailed instructions to carry out complex tasks involving unfamiliar process or procedures
- understand a wide range of recorded, broadcast and telephone audio material, which includes some non-standard usage, colloquialisms and regional accents
- identify finer points of detail including implicit attitudes and relationships between speakers

Speaking

Can use appropriate tone and suitable vocabulary when interacting with:

- Patients: adults/children/elderly/disabled/brain damaged/otherwise impaired/drunk
- Families of patients/carers
- Other doctors /senior and junior colleagues/GPs
- Nurses/Midwives
- Social workers/child protection workers
- Physiotherapists/dieticians/language therapists/paramedics/radiographers/lab. technicians
- Police/solicitors/coroners
- Pharmacists/drug companies' representatives
- Management/Admin. staff

Can use appropriate intonation, stress and word choice to:

- Give detailed information at handovers
- Discuss and evaluate the nature and relative merits of particular choices of procedures, courses of action or care packages
- Give instructions on carrying out a series of complex professional procedures
- Frame critical remarks in such a way as to minimize any offence
- Be persuasive
- Deliver bad news
- Request/receive/give results
- Explain complex, technical findings using suitably non-technical words and phrases
- Express sympathy or condolences, enquire into the causes of unhappiness or sadness and offer comfort
- Summarize what the patient is saying
- Deal with emotive circumstances
- Justify what they do (everything to everyone)

Can:

- pronounce words in a way that is readily comprehensible to those familiar with standard forms of English
- produce clear, smoothly flowing, well-structured speech which helps the hearer to notice and remember significant points
- modify their speech to express degrees of commitment or hesitancy, confidence or uncertainty
- adjust their level of formality and style of speech to suit the social context, formal, informal or colloquial as appropriate
- use appropriate technical terminology when discussing their area of specialisation with other specialists

Appendix 2: Comments on the IELTS test

Reading

Nurses

I don't think it gives you enough information as to whether a doctor has sufficient reading skills to practice in this country because it's missing critical skills areas they need.

It's a good general test but there should be some tasks that are more geared to the skills doctors need.

It's very clearly structured and it doesn't test someone's ability to pull information from unstructured content which is something medics will have to read.

There's some structure to the questions and it doesn't tackle unstructured or informal information.

All the answers are there so if they have time to read it, they should get it all right.

This is a very good test to tell if someone can read but I think there's scope to make it more specific, to fine tune it.

It's quite difficult.

Some of the paragraphs are quite subjective to the person reading them.

The first part is just recognition of words so you can go back and find the answers.

Patients

I didn't find any of it relevant to practice like for a doctor.

It's basic and straightforward and it doesn't ask anything very much in depth.

It's not like a comprehensive range of the stuff they will be coming up against.

I think these questions are all knowledge based.

It's a good test of reading ability in general but it's not sufficient to tell us whether a doctor can read the sorts of things they need to read.

Doctors

It's very factual as opposed to nuanced.

It's a reasonably good way to assess reading in English.

I think they are very fair questions. Sometimes the answers are a little bit more subtle than they would be in a guideline or hospital policy.

The majority of the questions, I thought, were relatively factual if you read it in enough detail the answers are there and they could be picked out.

It's very formulaic, isn't it?

IELTS is not enough by itself.

I don’t think it’s testing what it needs to test. I want something that gives more information.

We’ve identified that this is very much looking at one particular aspect of reading and doctors need to have skills in a lot of different aspects of reading.

AHPs

It doesn’t tell us whether a doctor is going to be able to read all the things we determined they should be able to read.

It’s a good test of your interpretation of implied things and I think it tests your ability to weed things out of a text.

It’s a good test but there should also be an additional, career specific, test.

ROs

It’s not easy.

I think we all thought it was quite challenging and that is reassuring.

It doesn’t provide enough evidence as to the ability of a prospective doctor to read what is required in the job.

We expect people to be able to read scientific documents as part of their continuing medical education so I think if they can’t read this and make sense of it then they are not going to be able to read an article in a journal.

Writing

Nurses

What I liked about this task is that they were having to consider various arguments and put them together and then come up with an opinion.

I think it weeds out some of the poorer candidates so it's quite good from that point of view but it feels like it should be more.

It's colloquial and I don't want colloquial

It elicits the sort of language with the content and you can see how they think logically and put the points in clear order so it's actually an adequate task.

This doesn't tell us whether these people can write as doctors.

It's very structured, clear.

They have to create arguments they are being asked to create an argument for some situation which is not a real life one for them. It's hypothetical and I don't think it's the most effective tool your can use for assessing someone's ability to write.

It shows that they understand what is being asked of them.

It gives an overall idea of their level of English but not the sort of writing doctors predominantly need to do which is factual, not invented.

I think it's nice little paragraphs to write.

It's an adequate task to get information.

Patients

I think reading and writing should be together.

It makes sense to integrate a lot of skills together.

It doesn't actually really reflect the sorts of writing that we think doctors have to do so it's very difficult.

We need different, more specific evidence.

I don't think this task is particularly appropriate for determining whether their writing skills are adequate for a doctor.

These aren't the kinds of writing we expected doctors to be able to do.

It's very difficult to judge the content and cohesion and everything.

I don't think any of these people were doctors and I don't think we feel this is a suitable test for doctors.

Doctors

It's a reasonable task and gives some indication that the writing is adequate for a doctor.

It doesn’t give me enough information whether a doctor could write what they need to write.

It’s very hard to tell really to transfer what you’re being told from the way they have answered these questions, would they be able to write sufficiently good English in the medical context.

I think we need multiple different inputs.

This is not an appropriate test for doctors and shouldn’t be used.

I think the writing and the speaking are being artificially divided.

It’s very simplistic and very different from what doctors have to write.

This test doesn’t enable us to predict anything to do with doctors’ writing skills, the sorts of skills a doctor is really going to need.

I don’t think the test best represents what needs to be tested.

AHPs

It’s quite a good indicator of vocabulary and structure and grammar and logic so I’m happy with the task.

I think that as a task it’s not bad in that they have to look at a topic and present an argument in kind of like a logical, concise way, so I guess you could transfer that over to work you would have to do at work. So looking through documents and coming up with an argument and being able to write it down in a clear way that someone else can follow. So I guess the structure isn’t bad.

Listening

Nurses

I liked the fact that it was listening into a conversation and having to get information from that rather than being spoken to directly, which I can see being a particular skill you might need.

I think we are trying to suggest that there are a variety of skills that a doctor needs to possess and that these questions do seem to address the different areas. But the way the test is set out you might completely miss one area.

I don't think they match up with all the things we've said doctors need to be able to listen to.

It's a very flat test, it's very similar in format throughout all the recordings.

They speak very clearly, the pronunciation is very, very clear and it runs chronologically so if we are going to assess somebody's ability to listen to real people in real life, I don't think it's a particularly useful too to assess people against what we are actually expecting them to do.

It seems very basic, it's dated in its format it's the same sort of stuff they were using for languages at school 20 years ago.

It's all actors makes it very stilted. There's no variety really in what we are asking them to do and the variety of listening they are going to have to do in their jobs isn't reflected in this.

The questions run chronologically which makes it very, very easy.

You could do very well on this test but how you function in an environment hasn't been tested at all.

Good, wide topics, different accents and different speeds. I think it's reasonable.

The accents are very broad and clear.

I think it's a reasonable test.

Patients

I don't think they should all be separate tests I think it makes sense to integrate the skills together.

I don't think it reflects what they are going to hear either so if they can't get the simple stuff on these tests they have no hope in a busy hospital really.

No way does it test the things we said earlier that doctors have to listen to.

I thought it was quite good when you had to extrapolate the information – it wasn't just stating, just repeating it.

This test won't give an indication of them (doctors) being able to do the things we think they should do.

I don’t think these tests are relevant to listening skills that a doctor would need. I think these are fine for somebody who wanted to come here to study but not in a busy situation that has background noise.

There certainly needs to be more diversity in the kinds of skills they are testing.

In some respects this is quite a good test of some of the things we think doctors should do but because of the tempo and the guidance that is given, if a doctor is going to be able to function in a genuine medical setting they would need to get all of the questions right.

This is too easy, basically. The listening seemed a level below the reading.

I like the fact that they used different accents and it was structured.

It’s not the sorts of things doctors would be listening to.

The telephone conversation was very clear compared to normal.

I think this is a bit too easy myself.

It’s an unnatural test in a way that it doesn’t represent what they will meet when they come here.

I don’t really think this test gives us an idea of what we want to see a doctor should be able to do.

Doctors

I like that the questions probe different memory systems. So there were clear semantic memory probes, word finding tasks. There were clear imagined memory word finding tasks and there were sort of connotative meaning, how it feels, so there was paraphrasing words which was more than imagined, there was word finding and there was another level where you had to understand the affect being expressed. And they used different styles of questions which I thought probed the different ways we render things rather than just relying on semantics or imagination.

I think the one thing it misses out on is understanding whether a person has got the whole gist of the situation.

I agree that this is a sufficient test and it tests your listening skills.

This audio is a lot easier to understand than a lot of hospitals, in a busy A&E and regional accents and colloquialisms, none of that. And people don’t speak in such neat sentences.

I think the questions are very appropriate.

Another way the test is imperfect apart from the telephone you don’t often listen to auditory without visual clues and I think that possibly counteracts the fact that although yes in a hospital there is more noise, people are talking on top of each other, you do have visual triggers.

I want the spoken to be less clear and quicker but I think the questions are very appropriate.

For me this is the right standard.

The accents weren’t particularly complicated.

All this test seems to be doing is writing down what you have heard.

I think part of the trouble is we are trying to assess each part of the language separately whereas they are all interlinked and you can't assess the full complexity of one aspect of it without assessing the other components at the same time.

It's a listening test but it's not at all appropriate as a measure for assessing whether a doctor can cope in a clinical setting of any description.

I don't think this test is appropriate at all and none of the skills are really tested.

The listening test is just farcical.

AHPs

I found it to be forced, unnatural conversation. It wasn't like natural conversation where it was flowing very quickly and you had to pick things up fast.

It needs something with background noise or distractions.

It's a good test of general listening ability but it's too clean, it's not listening in the real world and some of it is too artificial and structured.

Doctors need to be able to pick up information from much more messy discourse with background noise.

I think the listening was good because it involved a bit of writing so you had to listen and then write down points.

ROs

If they can't do this test, they won't be able to manage in A&E where they've got anxious relatives yelling at them and patients who are pretty incoherent.

I thought that the test was very clear and very fair and the structure of the questions was good.

They need to listen over the phone. Listening face-to-face and listening over the telephone are two important and sometimes different methods of communication.

The telephone is difficult, it's harder than face-to-face, isn't it. A telephone conversation with an agitated relative.

They need a test like the call handlers 999 ambulance services. Don't they have a test with agitated, stressed people? They have to be able, as part of their test, to get accurately location, symptoms, with an awful lot of background noise. One scenario they have got is a drunk person who is on the phone and they are tested on the accuracy of what they can get from the conversation.

Speaking

Nurses

I thought it was quite a good test, really. The good ones really stood out. I thought them having to present a topic was a good test of organizational ability in ordering your thoughts.

It's difficult to gauge whether someone is hesitating because they can't think of the right words or because they can't think of anything to say.

I think you need different topics because these were all relatively happy topics and you couldn't see how someone would show empathy breaking bad news.

The topics are quite immature, it's not topics adults would discuss with each other.

Getting your point across when there is more than one person talking across a cacophony of noise may be more challenging.

Patients

I think there should be an integrated test or some additional tasks to determine whether doctors could perform the sorts of things they need to do.

From the sorts of questions and the sort of tasks, we couldn't say whether someone could be a doctor.

It's all question and answer and it would have been better to give them a scenario to have a proper conversation.

Doctors

Part of the problem is we are using a test that was originally designed for one purpose for a completely different purpose so it's never going to be perfect.

I don't think it's broad enough and at the same time I don't think it's specific enough to medicine to provide the answers that as a doctor we would expect from our colleagues.

AHPs

I think it's good. I think it gives you a good idea of their fluidity of speaking, how they structure their responses and their sentences and their grammar and their vocabulary.

It's quite a good test because there is enough opportunity to display your language ability.

IELTS Overall

Nurses

I think the test is limited.

I think if you are stuck with this test and that is the only test that you have then the score of 8 would be minimum. But I think you can design a test that has much greater scope for what you are trying to get from this.

If you are saying that this is a test that we are going to use to see whether an individual has the communication skills in whatever format to function as a member of a medical team, then you could get much more useful information than from that. This is way too broad.

This test is not suited to the needs of what they are working in.

It's fine for an English test, you know perfect for school A level and maybe even for university but for somebody coming to work here, work in hospitals, work with people and ultimately making decisions on peoples' lives, this ….. doesn't work.

It's a good indicator of whether somebody has sufficient English to come to this country and apply for the PLAB test.

I think it's a starting point.

I think if it's challenged us as English speakers, then it certainly challenges somebody whose first language isn't English.

And I think there is a wide subject matter covered so I think it's a good reflection.

Patients

I don't think you could really show how they would survive in any kind of working environment.

I think there should be something more. I think we all feel there ought to be additional tests.

I mean, these test are not adequate.

We don't feel the listening test is appropriate at all which is why we set it so high and we don't actually feel that any of the skills tests are really appropriate.

I think the reading test was much more realistic than the writing test and as judges I felt we could say it (*the reading test*) was quite a good test for doctors but reading these, I don't think we could feel that they were doctors or that this (*the writing test*) was a suitable test for doctors.

It's a good test for someone coming to try to get into a British university but it doesn't actually test what we think a doctor should be able to do.

There should also be an integrated test or possibly some additional tasks to determine whether doctors could perform the sorts of things that doctors need to do in each of the skills.

Doctors

I think this is a screening test.

I wonder whether you can do the four separate tests and then add an extra test that combines the skills that are specific for doctors. That could be the addition, a couple of scenarios that utilize different skills.

Very often you have people coming from abroad and their technical knowledge and their ability to talk in technical terms is pretty decent. It's the getting into the local language and breaking down into simpler language that's the problem.

This is a general test for professionals on a wide scale and not necessarily aimed purely at doctors.

I think you won't be able to test the medical side of things unless you actually start testing specifically the can do statements. Because at the start we said that we think these are the best set we've had so if you design a test to address those statements, that might give you a better idea of someone's capability for functioning as a doctor.

It's almost a focusing test to see whether they should then proceed to the real doctoring test.

There is certainly benefit from having a significantly higher bar to pass in that if nothing else it saves the IMGs the expense of sitting the PLAB test if they are not going to get through this initial screening.

Maybe you have to look at adjusting the style; the listening test is just farcical.

One of the reasons for having this discussion is, is the test fit for the purposes we are looking for, and you know, I think clearly we are suggesting that it is not.

In my opinion there needs to be another test, or the PLAB needs to be changed and have a better focus on it, because I don't think this is going to give you enough information as it is about how someone is as a doctor.

I don't think it's broad enough and at the same time I don't think it's specific enough to provide the answers that as a doctor we would look for what we expect from our colleagues.

I can't say anything other than this structure of test for a potential doctor in this country is not appropriate.

I think in practical things that could be done, they could have a telephone consultation, just things like that, that would be a way of having a test that examines all the various areas of communication skills that we have been looking at today in an integrated fashion and in a context that is relevant to the job that they are going to be doing in this country,

AHPs

I think they could stand to do this test of their understanding and comprehension and general English skills but that there should also be an additional, career specific, test.

I was thinking that they could do reading and speaking and the speaking task should involve something written.

Yes, because I think if a doctor gets results or gets a written report from an x-ray or something, they have to read that and then go on to explain it in non-technical terms to the patient.

That's quite a high skill that they have to read the information, extract it and then be able to explain it in simple terms to the patient.

In the test as it stands you are prompted to look for certain answers but if you are given a script and you have to extract it yourself, that is a whole other skill.

ROs

We have spoken about putting a slightly more difficult element into it, listening, just to make it more representative of the direction that professionally….

Listening in stressful situations not necessarily as the whole bit but a section where you have to listen in a stressful situation.

Final Panel

Reading and Listening

I wonder looking at us, I think we probably got more confident in our decision making as the process went on and actually I would be inclined to think that our 7 for reading was probably a little on the lenient side, whereas towards the end our figures were much more 7.5, 8.5, so I wonder whether that skewed things slightly for our group in particular for the reading score.

The reading side of things will in many ways be more technical reading reports and things but there will also be written communication from the patients and it's important that IMG can read it, comprehend it, pick out the nuances and my inclination would be that going up to 8 is probably not a bad things because from what I remember of the test it was very much a kind of comprehension test. It seemed like the kind of thing you do at GCSE.

I would say as well with the reading it's all well and good to think with reading, OK you have plenty of time, but actually, realistically, in a busy medical setting you don't have a lot of time and I think it's taking that into consideration as well, particularly if you are on a round or about to see someone and you have who knows how many patients to get through, then you have to be able to scan through a set of notes, pick out the key points in the history, think about that, interpret it and then go and make your decision and further assessment from that. So I think it's all well and good to sit for half an hour reading something but actually, you don't realistically have the time.

I do remember thinking for the reading that it was a structured thing and if you had enough time to do it you could probably have picked out a key word that could then help you guess the answer without actually necessarily understanding the whole text. I think it you had enough time you could probably fumble your way through it without a true understanding.

And in terms of practical application as well there is a very big difference between sitting down and doing the test in a formal environment than to actually having to read something on the fly. Having a busy day in an emergency room is totally different, it's not like they can sit down in a nice quiet area and think about it for 20 minutes.

With really awful handwriting you have to read around the text and guess sometimes what a whole paragraph says and I think if you don't have a really good understanding of the writing around the particular word that you are trying to work out what it is, that's more advanced reading skills I suppose.

In a busy environment and in A & E, having to read something quickly, you are never going to be able to test that short of creating a kind of artificial situation where you are in a fake A & E with everything going on.

So in essence you are saying that just getting a standard reading test which is not medically specific. The higher the skills the better the chances will be when it comes to whatever you are reading. So we are saying that we want really good skills for reading.

If you want somebody who can read something in an environment where there are distractions you have to play something in the background that is going to distract them.

I think if we were to recommend that it was raised, I suppose you would hope that if you have a higher level on the test then that is going to…, like when you are in a more demanding, stressful situation, that somebody who is perhaps just scaring in on a 7 or 7.5 who may do OK on the test but actually when they are put into that more demanding environment may struggle more than someone who comes in with a higher basic level. So I would think it makes sense to raise it.

I think the reading test is probably a better test than the listening test because I don't think the listening test really picks up on any of the 'can do's.

Everybody thinks that the reading was much more difficult than the listening, that it's just on a different level altogether.

Well the listening was just a joke to my mind. It was just sitting down listening to what was said and writing down what was said in a box. It didn't show any need to comprehend or understand what has been said and interpret that, which as a doctor is what you have to do. You have to listen to what someone is saying to you, pick up on any undercurrents and hidden agendas and all the stuff they teach you in GP and then be able to use the information that you've got. It's not just a case of someone saying "I have this, this, this" and writing down "that, that, that and that".

It isn't really just about comprehension… and you know when you're seeing a patient, it's so much more than someone listing a set of symptoms, a lot of subtlety and meaning can be conveyed in the tone of voice but from what I remember, the listening test didn't really address that.

One of the things I think we agreed was that the listening skills need to be higher because listening is something you probably only get a chance to do once. The others you are more likely to be able to repeat other than in an emergency situation. But listening, people have a tendency to want to be listened to the first time.

You can ask perhaps once to have someone say something again but actually more than that it gets irritating and you lose the flow of the conversation.

And in terms of patients having confidence in their doctor, I think if you told them something and then they said "sorry, can you say that again…"

I think also about confidence in your doctor, but a lot of patients wouldn't have the confidence to say "did you understand what I am saying?" or "have you heard what I am saying?" because the power position is so different. People just feel intimidated quite often so they will come out saying "you didn't understand a word I said" but in the situation not feel able or confident to challenge it.

In terms of safety as well I would want to know that if somebody had been given an instruction to do something, that they understood it correctly and were then going to go off and do it properly and hadn't sort of thought I heard something, jot it down, and then gone off and carried out what they thought were the orders. That is frightening.

It's interesting because the general feeling is that the listening test was probably the most important and it's actually the least appropriate test, basic test and with that in mind….

It was really measured and quite slow and structured and bland accents.

That might be why we all scored it higher because we felt maybe that it wasn't discriminatory enough.

And that's why we put it so high, we didn't think it would discriminate to any real degree unless you had it at the highest level.

And I think listening to someone effectively, it's not just hearing it, it's listening to it and then responding to it as well and that in itself is quite a complex skill, to be able to reflect to someone that you have actually heard them correctly and that you are clear on what they are telling you.

It's not just about saying back what you heard.

It's about interpreting. I guess that goes on to speaking in that if you get the tone right you can judge the mood and be empathetic with people too.

The test itself doesn't actually test any kind of active listening and interpretation in that way and I think from that point of view we should set a high minimum.

I think it's clear from all of the listening that 8.5, that is the only one that is a definitive agreement, isn't it, between all of the groups that listening is the key and I do agree with that. Most of communication is about listening and taking in the information and particularly in that patient setting.

Writing and Speaking

The writing is a definitive agreement, isn't it, between all the groups.

We scored the speaking much lower than everyone else had. We said 7 was acceptable. In my area we recruit a lot of IMGs and we wondered actually if our experience with working with IMGs is affecting our perception of what you need to be able to do for the test. We are hearing somebody at Band 7 on this test and we are saying "we know people like that" and maybe that was affecting our scoring of it.

Could that be the other side of the coin, the fact that you have worked with IMGs and you know that practicing all these skills in a medical environment will improve it? Which is great but I think as a patient you still want a minimum level of competency.

If you meet that doctor in the first week of the four years as a patient, the fact that they are going to be better in 3 years' time is a problem, isn't it?

A lot of people's experience is with a junior doctor. That is who you see. So I think that the level should be at a higher level.

It was felt on balance by all the panels that they judged a 7.5 to be adequate but the actual candidates they listened to, they only deemed the ones who got Band 8 to be successful. So in other words, there is a conflict between the ways the actual marker was felt to be acceptable versus the candidates that were deemed to be successful. Personally I would say there should be a minimum score of Band 8.

Well, Band 8 was universally deemed as being adequate, the 7.5 got a mixed response.

Yes, the mean scores indicate 7.5 to be adequate but actually if you are looking down at how the information was derived, the anomaly is that everyone felt that Band 8 was deemed to be the minimum in terms of the actual candidates. But for some reason, the average score worked out at 7.5.

Doctors seem to have lower scores for speaking. Was there a particular outlier who made that score so low and artificially lowered the mean?

That is exactly what happened. There was a very senior doctor on one panel and a very and a very junior doctor on another panel who absolutely insisted on recording 6.5 which, as you say, artificially lowered the overall mean.

Which is why the score for doctors is 7.

So that's why it might have come out at 7.5 but the justification, if you like, for changing it to a Band 8 is probably best put that actually we believed the patients' view was more important than the professionals' view and therefore because the patients wanted Band 8 for speaking, that's why we said it. Justifying it by saying that actually the patients would feel more comfortable.

Appendix 3: Comments on the IMG-EEA distinction

If evidence of English language competence should ever be required of EEAs in order for them to be admitted to the GMC register, should they provide the same evidence as IMGs?

Nurses

Yes, they should all provide the same evidence.

It's a standard that you would expect from a performance point of view irrespective of where they come from. So yes, the standard should be the same.

I think they should all be the same.

Just because you're from a European country as opposed to elsewhere doesn't necessarily make your language skills adequate.

At the end of the day it's English literacy and the understanding of it.

Anybody who is not a native speaker of English should provide the same standard of evidence of language ability.

We should ask EEA doctors for the same evidence of language ability that we ask of IMG doctors.

Patients

It's only fair if they come to work in this country that people should be able to communicate in the language of the country.

It's a very serious issue, it's life or death.

Everyone should provide the same evidence – it's ridiculous not to.

Doctors

This is fundamentally about patient safety so despite the extra cost it might incur really everyone should undergo a test and they should go through the same rigorous test.

We all agree that there is a minimum standard of English that's needed to be a doctor.

I think anyone whose first language is not the language they are going to be conducting their job in should have to exhibit a proficiency in that second language.

I don't think it matters which country you come from that doesn't affect your ability to communicate in the language of the country you are going to .

Just because employment law in the EEA means that people can move freely and work freely doesn't lessen the language barriers that they are going to face so I think I agree there should be the same requirements.

And even if you did have some basic communication skills that you could use to travel in the area it's not ever necessarily enough to talk in medical language in the country that you are going to work in so it's worth demonstrating that you could.

The bottom line is patients' safety and we can't ensure patients' safety unless we ensure communication.

The fact that you have a medical degree doesn't mean to say that you are a safe doctor - part of being a safe doctor is being able to communicate.

Everybody should provide the same evidence of language ability, it's ridiculous not to.

It's only reasonable that both IMG and EEA, rest of the world and Europeans should all provide the same evidence of ability to speak English.

AHPs

I think everybody should be made to provide evidence. I don't see how you can differentiate in any way. You can have someone from a far away country who speaks no English at all or speaks brilliant English or you can get someone from France who speaks none. If you have a test for everyone to prove that they can speak English that test should be for anyone for whom English is the second language.

I think it you have done your degree in English as your first language then perhaps you can have that waived but not if you have done university in another country.

ROs

There should be no difference between a graduate from the UK or an IMG or a EEA doctor.

The issue is more about which language you are educated in isn't it. If you are educated in English even in France you could argue that the person's language skills ought to be adequate.

Language skills are colloquial, it's all about how you use the language, language skills are about the use of language particularly about the spoken language.

Consequences are much greater than going to the supermarket and ending up with the wrong vegetables.

If reception of what the patients are saying is misinterpreted it can go in all sorts of directions that aren't appropriate.

There is a separate argument about whether UK graduates should be subjected to the same test.

You can come in as a member of your family and become an undergraduate and not have the language skills. That is an entry to your undergraduate career. We should start out with that base. I think it is important that the issue about UK graduates is also considered.

If there is a principle of fairness then the argument to test everyone is the right one to take.

Our first duty has to be to the patients.

Doctors have got to have the ability to communicate at the level of the patients on a level that the patients can actually understand.

They should decide to put English or language testing into medical school curriculum that would mean you couldn't graduate without achieving a standard of communication.

If there is a sort of escape route for IMGs to become EEA then that is an issue so all IMG and EEA graduates should provide the same evidence of language ability.

Final panel

Would an RO be more conscious of the linguistic skills of a foreign doctor than they would for a native?

I think they will apply the same standard to everybody; there is no potential for being specific. You don't necessarily know just by looking at somebody or by reading somebody's name where they qualified from anyway. There are an awful lot of British graduates who don't have traditionally Anglo-Saxon names.

The GMC is responsible for fitness to practise and licence to practise and if you can make it a requirement of registration that there is a demonstration of competence in English and you can probably specify GCSE grades or IELTS score for people who don't have that GCSE, by making it a fitness to practise issue you can get round EU law that way but only if it's applied to everyone.

That's why the RO and medical directors' panel would have no problem applying it to everyone.

The bottom line is that IMGs will be asked to take the IELTS test with the previously specified banding levels requirements and then go on to take the PLAB and then to enter revalidation. The EEA doctors and those who were previously IMGs but became EEA doctors because of EC rights, we are asking that the GMC will take on a function to make it a mandatory requirement that somehow they provide alternative proof to IELTS at Band 8.

Ultimately this is all about patients' safety and it's a step in the right direction if any doctor who doesn't have English as their first language, if they have to take an English language test, that is at least a step in the right direction in trying to make things safer for patients. And if by introducing IELTS for EEA doctors screens out a handful who otherwise would have had poor communication and made mistakes as a result of that, and it means a handful of fewer problems to patients, that is a good thing. If that's what comes out of this panel then that is fantastic. And even if all we do is recommend IELTS for EEA doctors as a demonstration of English language, it's a small step but at least it's a step in the right direction.

Appendix 4: Analysis of Variance (ANOVA) of reading and listening scores

By PANEL

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Reading	Between Groups	504.863	10	50.486	5.826	.000
	Within Groups	441.976	51	8.666		
	Total	946.839	61			
Listening	Between Groups	218.495	10	21.850	6.733	.000
	Within Groups	165.505	51	3.245		
	Total	384.000	61			

Multiple Comparisons

Scheffe

Dependent Variable	(I) Panel	(J) Panel	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Reading	LP	NP	.02381	1.63780	1.000	-7.3412	7.3888
		CP	1.88095	1.63780	.999	-5.4840	9.2459
		EAN	-1.43333	1.78258	1.000	-9.4494	6.5827
		BN	-4.83333	1.78258	.688	-12.8494	3.1827
		LN	-4.23333	1.78258	.835	-12.2494	3.7827
		EAD	3.76667	1.78258	.916	-4.2494	11.7827
		LD	2.16667	1.78258	.999	-5.8494	10.1827

		DD	.76667	1.78258	1.000	-7.2494	8.7827
		AHP	4.56667	1.78258	.759	-3.4494	12.5827
		RO-MD	-2.97619	1.63780	.969	-10.3412	4.3888
	NP	LP	-.02381	1.63780	1.000	-7.3888	7.3412
		CP	1.85714	1.57355	.999	-5.2189	8.9332
		EAN	-1.45714	1.72374	1.000	-9.2085	6.2943
		BN	-4.85714	1.72374	.635	-12.6085	2.8943
		LN	-4.25714	1.72374	.798	-12.0085	3.4943
		EAD	3.74286	1.72374	.901	-4.0085	11.4943
		LD	2.14286	1.72374	.998	-5.6085	9.8943
		DD	.74286	1.72374	1.000	-7.0085	8.4943
		AHP	4.54286	1.72374	.725	-3.2085	12.2943
		RO-MD	-3.00000	1.57355	.957	-10.0760	4.0760
	CP	LP	-1.88095	1.63780	.999	-9.2459	5.4840
		NP	-1.85714	1.57355	.999	-8.9332	5.2189
		EAN	-3.31429	1.72374	.954	-11.0657	4.4371
		BN	-6.71429	1.72374	.160	-14.4657	1.0371
		LN	-6.11429	1.72374	.279	-13.8657	1.6371
		EAD	1.88571	1.72374	.999	-5.8657	9.6371
		LD	.28571	1.72374	1.000	-7.4657	8.0371
		DD	-1.11429	1.72374	1.000	-8.8657	6.6371
		AHP	2.68571	1.72374	.990	-5.0657	10.4371
		RO-MD	-4.85714	1.57355	.495	-11.9332	2.2189

	EAN	LP	1.43333	1.78258	1.000	-6.5827	9.4494
		NP	1.45714	1.72374	1.000	-6.2943	9.2085
		CP	3.31429	1.72374	.954	-4.4371	11.0657
		BN	-3.40000	1.86185	.968	-11.7725	4.9725
		LN	-2.80000	1.86185	.992	-11.1725	5.5725
		EAD	5.20000	1.86185	.647	-3.1725	13.5725
		LD	3.60000	1.86185	.952	-4.7725	11.9725
		DD	2.20000	1.86185	.999	-6.1725	10.5725
		AHP	6.00000	1.86185	.426	-2.3725	14.3725
		RO-MD	-1.54286	1.72374	1.000	-9.2943	6.2085
	BN	LP	4.83333	1.78258	.688	-3.1827	12.8494
		NP	4.85714	1.72374	.635	-2.8943	12.6085
		CP	6.71429	1.72374	.160	-1.0371	14.4657
		EAN	3.40000	1.86185	.968	-4.9725	11.7725
		LN	.60000	1.86185	1.000	-7.7725	8.9725
		EAD	8.60000*	1.86185	.038	.2275	16.9725
		LD	7.00000	1.86185	.201	-1.3725	15.3725
		DD	5.60000	1.86185	.536	-2.7725	13.9725
		AHP	9.40000*	1.86185	.014	1.0275	17.7725
		RO-MD	1.85714	1.72374	1.000	-5.8943	9.6085
	LN	LP	4.23333	1.78258	.835	-3.7827	12.2494
		NP	4.25714	1.72374	.798	-3.4943	12.0085
		CP	6.11429	1.72374	.279	-1.6371	13.8657

		EAN	2.80000	1.86185	.992	-5.5725	11.1725
		BN	-.60000	1.86185	1.000	-8.9725	7.7725
		EAD	8.00000	1.86185	.076	-.3725	16.3725
		LD	6.40000	1.86185	.325	-1.9725	14.7725
		DD	5.00000	1.86185	.701	-3.3725	13.3725
		AHP	8.80000*	1.86185	.030	.4275	17.1725
		RO-MD	1.25714	1.72374	1.000	-6.4943	9.0085
	EAD	LP	-3.76667	1.78258	.916	-11.7827	4.2494
		NP	-3.74286	1.72374	.901	-11.4943	4.0085
		CP	-1.88571	1.72374	.999	-9.6371	5.8657
		EAN	-5.20000	1.86185	.647	-13.5725	3.1725
		BN	-8.60000*	1.86185	.038	-16.9725	-.2275
		LN	-8.00000	1.86185	.076	-16.3725	.3725
		LD	-1.60000	1.86185	1.000	-9.9725	6.7725
		DD	-3.00000	1.86185	.987	-11.3725	5.3725
		AHP	.80000	1.86185	1.000	-7.5725	9.1725
		RO-MD	-6.74286	1.72374	.156	-14.4943	1.0085
	LD	LP	-2.16667	1.78258	.999	-10.1827	5.8494
		NP	-2.14286	1.72374	.998	-9.8943	5.6085
		CP	-.28571	1.72374	1.000	-8.0371	7.4657
		EAN	-3.60000	1.86185	.952	-11.9725	4.7725
		BN	-7.00000	1.86185	.201	-15.3725	1.3725
		LN	-6.40000	1.86185	.325	-14.7725	1.9725

		EAD	1.60000	1.86185	1.000	-6.7725	9.9725
		DD	-1.40000	1.86185	1.000	-9.7725	6.9725
		AHP	2.40000	1.86185	.998	-5.9725	10.7725
		RO-MD	-5.14286	1.72374	.549	-12.8943	2.6085
	DD	LP	-.76667	1.78258	1.000	-8.7827	7.2494
		NP	-.74286	1.72374	1.000	-8.4943	7.0085
		CP	1.11429	1.72374	1.000	-6.6371	8.8657
		EAN	-2.20000	1.86185	.999	-10.5725	6.1725
		BN	-5.60000	1.86185	.536	-13.9725	2.7725
		LN	-5.00000	1.86185	.701	-13.3725	3.3725
		EAD	3.00000	1.86185	.987	-5.3725	11.3725
		LD	1.40000	1.86185	1.000	-6.9725	9.7725
		AHP	3.80000	1.86185	.932	-4.5725	12.1725
		RO-MD	-3.74286	1.72374	.901	-11.4943	4.0085
	AHP	LP	-4.56667	1.78258	.759	-12.5827	3.4494
		NP	-4.54286	1.72374	.725	-12.2943	3.2085
		CP	-2.68571	1.72374	.990	-10.4371	5.0657
		EAN	-6.00000	1.86185	.426	-14.3725	2.3725
		BN	-9.40000*	1.86185	.014	-17.7725	-1.0275
		LN	-8.80000*	1.86185	.030	-17.1725	-.4275
		EAD	-.80000	1.86185	1.000	-9.1725	7.5725
		LD	-2.40000	1.86185	.998	-10.7725	5.9725
		DD	-3.80000	1.86185	.932	-12.1725	4.5725

		RO-MD	-7.54286	1.72374	.064	-15.2943	.2085
	RO-MD	LP	2.97619	1.63780	.969	-4.3888	10.3412
		NP	3.00000	1.57355	.957	-4.0760	10.0760
		CP	4.85714	1.57355	.495	-2.2189	11.9332
		EAN	1.54286	1.72374	1.000	-6.2085	9.2943
		BN	-1.85714	1.72374	1.000	-9.6085	5.8943
		LN	-1.25714	1.72374	1.000	-9.0085	6.4943
		EAD	6.74286	1.72374	.156	-1.0085	14.4943
		LD	5.14286	1.72374	.549	-2.6085	12.8943
		DD	3.74286	1.72374	.901	-4.0085	11.4943
		AHP	7.54286	1.72374	.064	-.2085	15.2943
Listening	LP	NP	.80952	1.00223	1.000	-3.6974	5.3164
		CP	.23810	1.00223	1.000	-4.2688	4.7450
		EAN	-.53333	1.09083	1.000	-5.4386	4.3720
		BN	-.53333	1.09083	1.000	-5.4386	4.3720
		LN	-2.73333	1.09083	.783	-7.6386	2.1720
		EAD	4.66667	1.09083	.079	-.2386	9.5720
		LD	-.33333	1.09083	1.000	-5.2386	4.5720
		DD	-2.53333	1.09083	.854	-7.4386	2.3720
		AHP	-.33333	1.09083	1.000	-5.2386	4.5720
		RO-MD	-2.33333	1.00223	.852	-6.8402	2.1736
	NP	LP	-.80952	1.00223	1.000	-5.3164	3.6974
		CP	-.57143	.96291	1.000	-4.9015	3.7586

		EAN	-1.34286	1.05482	.998	-6.0862	3.4005
		BN	-1.34286	1.05482	.998	-6.0862	3.4005
		LN	-3.54286	1.05482	.360	-8.2862	1.2005
		EAD	3.85714	1.05482	.237	-.8862	8.6005
		LD	-1.14286	1.05482	1.000	-5.8862	3.6005
		DD	-3.34286	1.05482	.453	-8.0862	1.4005
		AHP	-1.14286	1.05482	1.000	-5.8862	3.6005
		RO-MD	-3.14286	.96291	.405	-7.4729	1.1872
	CP	LP	-.23810	1.00223	1.000	-4.7450	4.2688
		NP	.57143	.96291	1.000	-3.7586	4.9015
		EAN	-.77143	1.05482	1.000	-5.5148	3.9719
		BN	-.77143	1.05482	1.000	-5.5148	3.9719
		LN	-2.97143	1.05482	.635	-7.7148	1.7719
		EAD	4.42857	1.05482	.092	-.3148	9.1719
		LD	-.57143	1.05482	1.000	-5.3148	4.1719
		DD	-2.77143	1.05482	.729	-7.5148	1.9719
		AHP	-.57143	1.05482	1.000	-5.3148	4.1719
		RO-MD	-2.57143	.96291	.708	-6.9015	1.7586
	EAN	LP	.53333	1.09083	1.000	-4.3720	5.4386
		NP	1.34286	1.05482	.998	-3.4005	6.0862
		CP	.77143	1.05482	1.000	-3.9719	5.5148
		BN	.00000	1.13933	1.000	-5.1234	5.1234
		LN	-2.20000	1.13933	.953	-7.3234	2.9234

		EAD	5.20000*	1.13933	.043	.0766	10.3234
		LD	.20000	1.13933	1.000	-4.9234	5.3234
		DD	-2.00000	1.13933	.976	-7.1234	3.1234
		AHP	.20000	1.13933	1.000	-4.9234	5.3234
		RO-MD	-1.80000	1.05482	.980	-6.5434	2.9434
	BN	LP	.53333	1.09083	1.000	-4.3720	5.4386
		NP	1.34286	1.05482	.998	-3.4005	6.0862
		CP	.77143	1.05482	1.000	-3.9719	5.5148
		EAN	.00000	1.13933	1.000	-5.1234	5.1234
		LN	-2.20000	1.13933	.953	-7.3234	2.9234
		EAD	5.20000*	1.13933	.043	.0766	10.3234
		LD	.20000	1.13933	1.000	-4.9234	5.3234
		DD	-2.00000	1.13933	.976	-7.1234	3.1234
		AHP	.20000	1.13933	1.000	-4.9234	5.3234
		RO-MD	-1.80000	1.05482	.980	-6.5434	2.9434
	LN	LP	2.73333	1.09083	.783	-2.1720	7.6386
		NP	3.54286	1.05482	.360	-1.2005	8.2862
		CP	2.97143	1.05482	.635	-1.7719	7.7148
		EAN	2.20000	1.13933	.953	-2.9234	7.3234
		BN	2.20000	1.13933	.953	-2.9234	7.3234
		EAD	7.40000*	1.13933	.000	2.2766	12.5234
		LD	2.40000	1.13933	.917	-2.7234	7.5234
		DD	.20000	1.13933	1.000	-4.9234	5.3234

		AHP	2.40000	1.13933	.917	-2.7234	7.5234
		RO-MD	.40000	1.05482	1.000	-4.3434	5.1434
	EAD	LP	-4.66667	1.09083	.079	-9.5720	.2386
		NP	-3.85714	1.05482	.237	-8.6005	.8862
		CP	-4.42857	1.05482	.092	-9.1719	.3148
		EAN	-5.20000*	1.13933	.043	-10.3234	-.0766
		BN	-5.20000*	1.13933	.043	-10.3234	-.0766
		LN	-7.40000*	1.13933	.000	-12.5234	-2.2766
		LD	-5.00000	1.13933	.063	-10.1234	.1234
		DD	-7.20000*	1.13933	.000	-12.3234	-2.0766
		AHP	-5.00000	1.13933	.063	-10.1234	.1234
		RO-MD	-7.00000*	1.05482	.000	-11.7434	-2.2566
	LD	LP	.33333	1.09083	1.000	-4.5720	5.2386
		NP	1.14286	1.05482	1.000	-3.6005	5.8862
		CP	.57143	1.05482	1.000	-4.1719	5.3148
		EAN	-.20000	1.13933	1.000	-5.3234	4.9234
		BN	-.20000	1.13933	1.000	-5.3234	4.9234
		LN	-2.40000	1.13933	.917	-7.5234	2.7234
		EAD	5.00000	1.13933	.063	-.1234	10.1234
		DD	-2.20000	1.13933	.953	-7.3234	2.9234
		AHP	.00000	1.13933	1.000	-5.1234	5.1234
		RO-MD	-2.00000	1.05482	.958	-6.7434	2.7434
	DD	LP	2.53333	1.09083	.854	-2.3720	7.4386

		NP	3.34286	1.05482	.453	-1.4005	8.0862
		CP	2.77143	1.05482	.729	-1.9719	7.5148
		EAN	2.00000	1.13933	.976	-3.1234	7.1234
		BN	2.00000	1.13933	.976	-3.1234	7.1234
		LN	-.20000	1.13933	1.000	-5.3234	4.9234
		EAD	7.20000*	1.13933	.000	2.0766	12.3234
		LD	2.20000	1.13933	.953	-2.9234	7.3234
		AHP	2.20000	1.13933	.953	-2.9234	7.3234
		RO-MD	.20000	1.05482	1.000	-4.5434	4.9434
	AHP	LP	.33333	1.09083	1.000	-4.5720	5.2386
		NP	1.14286	1.05482	1.000	-3.6005	5.8862
		CP	.57143	1.05482	1.000	-4.1719	5.3148
		EAN	-.20000	1.13933	1.000	-5.3234	4.9234
		BN	-.20000	1.13933	1.000	-5.3234	4.9234
		LN	-2.40000	1.13933	.917	-7.5234	2.7234
		EAD	5.00000	1.13933	.063	-.1234	10.1234
		LD	.00000	1.13933	1.000	-5.1234	5.1234
		DD	-2.20000	1.13933	.953	-7.3234	2.9234
		RO-MD	-2.00000	1.05482	.958	-6.7434	2.7434
	RO-MD	LP	2.33333	1.00223	.852	-2.1736	6.8402
		NP	3.14286	.96291	.405	-1.1872	7.4729
		CP	2.57143	.96291	.708	-1.7586	6.9015
		EAN	1.80000	1.05482	.980	-2.9434	6.5434

	BN	1.80000	1.05482	.980	-2.9434	6.5434
	LN	-.40000	1.05482	1.000	-5.1434	4.3434
	EAD	7.00000*	1.05482	.000	2.2566	11.7434
	LD	2.00000	1.05482	.958	-2.7434	6.7434
	DD	-.20000	1.05482	1.000	-4.9434	4.5434
	AHP	2.00000	1.05482	.958	-2.7434	6.7434

*. The mean difference is significant at the 0.05 level.

By CATEGORY

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Reading	Between Groups	433.515	4	108.379	12.034	.000
	Within Groups	513.324	57	9.006		
	Total	946.839	61			
Listening	Between Groups	63.933	4	15.983	2.846	.032
	Within Groups	320.067	57	5.615		
	Total	384.000	61			

Multiple Comparisons

Scheffe

Dependent Variable	(I) Category	(J) Category	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Reading	patients	nurses	-4.16667*	1.02502	.005	-7.4298	-.9036
		doctors	1.56667	1.02502	.675	-1.6964	4.8298
		AHP	3.90000	1.50047	.165	-.8767	8.6767
		RO-MD	-3.64286	1.31788	.121	-7.8383	.5525
	nurses	patients	4.16667*	1.02502	.005	.9036	7.4298
		doctors	5.73333*	1.09579	.000	2.2449	9.2217
		AHP	8.06667*	1.54968	.000	3.1333	13.0000
		RO-MD	.52381	1.37365	.997	-3.8491	4.8967
	doctors	patients	-1.56667	1.02502	.675	-4.8298	1.6964
		nurses	-5.73333*	1.09579	.000	-9.2217	-2.2449
		AHP	2.33333	1.54968	.688	-2.6000	7.2667
		RO-MD	-5.20952*	1.37365	.011	-9.5825	-.8366
	AHP	patients	-3.90000	1.50047	.165	-8.6767	.8767
		nurses	-8.06667*	1.54968	.000	-13.0000	-3.1333
		doctors	-2.33333	1.54968	.688	-7.2667	2.6000
		RO-MD	-7.54286*	1.75717	.003	-13.1367	-1.9490
	RO-MD	patients	3.64286	1.31788	.121	-.5525	7.8383
		nurses	-.52381	1.37365	.997	-4.8967	3.8491
		doctors	5.20952*	1.37365	.011	.8366	9.5825
		AHP	7.54286*	1.75717	.003	1.9490	13.1367

Listening	patients	nurses	-1.63333	.80939	.406	-4.2100	.9433
		doctors	.23333	.80939	.999	-2.3433	2.8100
		AHP	-.70000	1.18482	.986	-4.4718	3.0718
		RO-MD	-2.70000	1.04064	.167	-6.0128	.6128
	nurses	patients	1.63333	.80939	.406	-.9433	4.2100
		doctors	1.86667	.86527	.337	-.8879	4.6212
		AHP	.93333	1.22368	.964	-2.9622	4.8288
		RO-MD	-1.06667	1.08467	.913	-4.5197	2.3863
	doctors	patients	-.23333	.80939	.999	-2.8100	2.3433
		nurses	-1.86667	.86527	.337	-4.6212	.8879
		AHP	-.93333	1.22368	.964	-4.8288	2.9622
		RO-MD	-2.93333	1.08467	.136	-6.3863	.5197
	AHP	patients	.70000	1.18482	.986	-3.0718	4.4718
		nurses	-.93333	1.22368	.964	-4.8288	2.9622
		doctors	.93333	1.22368	.964	-2.9622	4.8288
		RO-MD	-2.00000	1.38752	.722	-6.4171	2.4171
	RO-MD	patients	2.70000	1.04064	.167	-.6128	6.0128
		nurses	1.06667	1.08467	.913	-2.3863	4.5197
		doctors	2.93333	1.08467	.136	-.5197	6.3863
		AHP	2.00000	1.38752	.722	-2.4171	6.4171

*. The mean difference is significant at the 0.05 level.

By GENDER

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Reading	Between Groups	1.153	1	1.153	.073	.788
	Within Groups	945.685	60	15.761		
	Total	946.839	61			
Listening	Between Groups	3.165	1	3.165	.499	.483
	Within Groups	380.835	60	6.347		
	Total	384.000	61			

By ETHNICITY

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Reading	Between Groups	40.538	1	40.538	2.684	.107
	Within Groups	906.301	60	15.105		
	Total	946.839	61			
Listening	Between Groups	1.292	1	1.292	.203	.654
	Within Groups	382.708	60	6.378		
	Total	384.000	61			

By AGE

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Reading	Between Groups	106.778	4	26.694	1.811	.139
	Within Groups	840.061	57	14.738		
	Total	946.839	61			
Listening	Between Groups	3.856	4	.964	.145	.965
	Within Groups	380.144	57	6.669		
	Total	384.000	61			

Multiple Comparisons

Scheffe

Dependent Variable	(I) age	(J) age	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Reading	20s	30s	-.58889	1.34213	.996	-4.8615	3.6837
		40s	-3.61667	1.48684	.221	-8.3499	1.1166
		50s	-2.03333	1.48684	.759	-6.7666	2.6999
		60+	-2.00000	1.98245	.906	-8.3110	4.3110
	30s	20s	.58889	1.34213	.996	-3.6837	4.8615
		40s	-3.02778	1.43071	.356	-7.5824	1.5268
		50s	-1.44444	1.43071	.906	-5.9990	3.1101
		60+	-1.41111	1.94071	.970	-7.5893	4.7670
	40s	20s	3.61667	1.48684	.221	-1.1166	8.3499
		30s	3.02778	1.43071	.356	-1.5268	7.5824

		50s	1.58333	1.56726	.905	-3.4060	6.5726
		60+	1.61667	2.04346	.959	-4.8886	8.1219
	50s	20s	2.03333	1.48684	.759	-2.6999	6.7666
		30s	1.44444	1.43071	.906	-3.1101	5.9990
		40s	-1.58333	1.56726	.905	-6.5726	3.4060
		60+	.03333	2.04346	1.000	-6.4719	6.5386
	60+	20s	2.00000	1.98245	.906	-4.3110	8.3110
		30s	1.41111	1.94071	.970	-4.7670	7.5893
		40s	-1.61667	2.04346	.959	-8.1219	4.8886
		50s	-.03333	2.04346	1.000	-6.5386	6.4719
Listening	20s	30s	-.27778	.90284	.999	-3.1519	2.5964
		40s	-.66667	1.00019	.978	-3.8507	2.5174
		50s	-.33333	1.00019	.998	-3.5174	2.8507
		60+	-.73333	1.33359	.989	-4.9787	3.5121
	30s	20s	.27778	.90284	.999	-2.5964	3.1519
		40s	-.38889	.96243	.997	-3.4527	2.6750
		50s	-.05556	.96243	1.000	-3.1194	3.0083
		60+	-.45556	1.30551	.998	-4.6116	3.7005
	40s	20s	.66667	1.00019	.978	-2.5174	3.8507
		30s	.38889	.96243	.997	-2.6750	3.4527
		50s	.33333	1.05429	.999	-3.0230	3.6896
		60+	-.06667	1.37463	1.000	-4.4427	4.3094
	50s	20s	.33333	1.00019	.998	-2.8507	3.5174

		30s	.05556	.96243	1.000	-3.0083	3.1194
		40s	-.33333	1.05429	.999	-3.6896	3.0230
		60+	-.40000	1.37463	.999	-4.7761	3.9761
	60+	20s	.73333	1.33359	.989	-3.5121	4.9787
		30s	.45556	1.30551	.998	-3.7005	4.6116
		40s	.06667	1.37463	1.000	-4.3094	4.4427
		50s	.40000	1.37463	.999	-3.9761	4.7761

Appendix 5a: Many Facet Rasch Output – Speaking

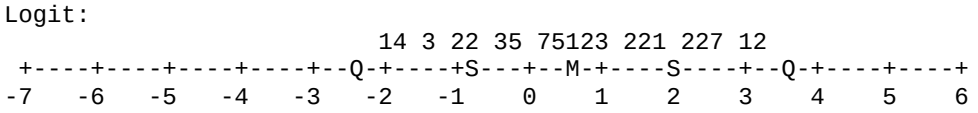
```
Facets (Many-Facet Rasch Measurement) Version No. 3.70.1  Copyright(c) 1987-
2012, John M. Linacre. All rights reserved.

22/10/2012 21:37:47

GMC 2012 Speaking Data 22/10/2012 21:37:47

Table 1. Specifications from file "C:\Users\Barry\Documents\CLARe\GMC
Project\gmcspeakdat2.txt".

Title = GMC 2012 Speaking Data 22/10/2012 21:37:47
Data file = (C:\Users\Barry\Documents\CLARe\GMC Project\gmcspeakdat2.txt)
Output file = gmcspeak.out

; Data specification
Facets = 3
Delements = N
Non-centered =  1
Positive =  1
Labels =
 1,Judges ; (elements = 57
 2,Testees ; (elements = 12
 3,Group ; (elements = 5)
Model = ?,?,,R5,1

; Output description
Arrange tables in order = N
Bias/Interaction direction = plus ; ability, easiness, leniency: higher score =
positive logit
Fair score = Mean
Pt-biserial = Measure
Heading lines in output data files = Y
Omit unobserved elements = yes
Barchart = Yes
Total score for elements = Yes
T3onscreen show only one line on screen iteration report = Y
T4MAX maximum number of unexpected observations reported in Table 4 = 100
T8NBC show table 8 numbers-barcharts-curves = NBC
Unexpected observations reported if standardized residual >= 3
Usort unexpected observations sort order = u
WHexact - Wilson-Hilferty standardization = Y

; Convergence control
Convergence = .5, .01
Iterations (maximum) = 0 ; unlimited
Xtreme scores adjusted by = .3, .5  ;(estimation, bias)

GMC 2012 Speaking Data 22/10/2012 21:37:47
Table 2. Data Summary Report.

Assigning models to Data= "C:\Users\Barry\Documents\CLARe\GMC
Project\gmcspeakdat2.txt"

In data: 20, 1, 1, 5
Check (1)? Unspecified element: facet: 1 element: 20 20 in line 110 - see Table
2. Delements = N
```

```
Check (2)? Invalid datum location: 19,2,1,6 in line 171. Datum "6" is too big
or not a positive integer, treated as missing.

Total lines in data file = 745

Total data lines = 744
Responses matched to model: ?,?,,R5,1 = 684
    Total non-blank responses found = 744
Responses with unspecified elements = 60
             Number of blank lines = 1
Number of invalid observations treated as missing = 1
Valid responses used for estimation = 683

List of unspecified elements. Please copy and paste into your specification
file, where needed

Labels=Nobuild ; to suppress this list
 1, Judges,  ; facet 1
 20 = 20
 32 = 32
 34 = 34
 46 = 46
 56 = 56
 *

GMC 2012 Speaking Data 22/10/2012 21:37:47
Table 3. Iteration Report.
+-------------------------------------------------------------+
| Iteration      Max. Score Residual      Max. Logit Change |
|             Elements    %  Categories   Elements    Steps |
|-------------------------------------------------------------|
| PROX   1                                  -3.9005          |
| PROX   2                                    .6200          |
| JMLE   3     29.0387  13.0  -120.3444       .5596   1.2513 |
| JMLE   4     12.9384   5.7   -19.2269      -.2653    .2030 |
| JMLE   5     11.2492   4.9   -18.4749      -.2195    .1668 |
| JMLE   6     10.2793   4.5   -16.3399      -.1862    .1387 |
| JMLE   7      9.4322   4.1   -14.6010      -.1605    .1185 |
*
| JMLE  74       .3581    .2     -.4711      -.0110    .0081 |
| JMLE  75       .3429    .2     -.4511      -.0107    .0079 |
| JMLE  76       .3283    .1     -.4318      -.0103    .0076 |
| JMLE  77       .3141    .1     -.4133      -.0100    .0073 |
+-------------------------------------------------------------+
Subset connection O.K.



GMC 2012 Speaking Data 22/10/2012 21:37:47
Table 4. Unexpected Responses - appears after Table 8.

GMC 2012 Speaking Data 22/10/2012 21:37:47
Table 5. Measurable Data Summary.

+-------------------------------------------------+
| Cat  Score  Exp.  Resd StRes|                    |
|-----------------------------+--------------------|
| 3.06  3.06  3.06   .00 -.05 | Mean (Count: 683)  |
| 1.54  1.54  1.44   .51 1.13 | S.D. (Population)  |
| 1.54  1.54  1.45   .51 1.13 | S.D. (Sample)      |
+-------------------------------------------------+
Data log-likelihood chi-square = 894.8979
Approximate degrees of freedom = 613
```

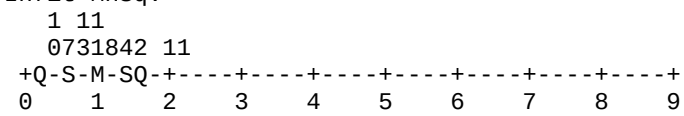
```
Chi-square significance prob.  = .0000
                                          Count   Mean   S.D.   Params
Responses used for estimation      =        683   3.06   1.54       70
Count of measurable responses      =        683
Raw-score variance of observations   =    2.37 100.00%
Variance explained by Rasch measures =    2.11  89.19%
Variance of residuals                =    0.26  10.81%


GMC 2012 Speaking Data 22/10/2012 21:37:47
Table 6.0  All Facet Vertical "Rulers".

Vertical = (1*,2A,3A,S) Yardstick (columns lines low high extreme)= 0,3,7,6,End
+------------------------------+
|Measr|+Judges |-Testees|Scale|
|-----+--------+--------+-----|
|   6 +        +        + (5) |
|     |        |        |     |
|     |        | 9      |     |
|   5 +        +        +     |
|     |        | 11     |     |
|     |        | 7      |     |
|   4 +        +        + ---- |
|     |        | 4      |     |
|     | *      |        |     |
|   3 + .      +        +     |
|     | ***.   |        |     |
|     | *      | 2      |  4  |
|   2 + *.     +        +     |
|     | *      | 6      |     |
|     | *      |        |     |
|   1 + *.     +        +     |
|     | *.     |        | --- |
|     | ****** | 10     |     |
*   0 * **.    *        *     *
|     | *.     |        |  3  |
|     | *      | 1      |     |
|  -1 + *      +        +     |
|     | *.     |        | --- |
|     | **     |        |     |
|  -2 + .      + 5      +     |
|     |        |        |  2  |
|     |        |        |     |
|  -3 +        +        +     |
|     |        |        | --- |
|     |        |        |     |
|  -4 +        +        +     |
|     |        |        |     |
|     |        |        |     |
|  -5 +        +        +     |
|     |        |        |     |
|     |        | 8      |     |
|  -6 +        +        +     |
|     |        |        |     |
|     |        |        |     |
|  -7 +        + 12  3  + (1) |
|-----+--------+--------+-----|
|Measr| * = 2  |-Testees|Scale|
+------------------------------+
S.1: Model = ?,?,,R5
```

GMC 2012 Speaking Data 22/10/2012 21:37:47
Table 6.1 Judges Facet Summary.

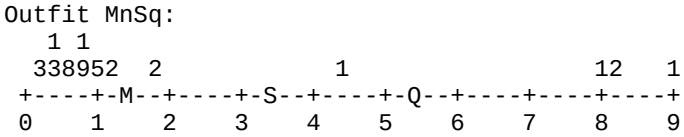
Logit:



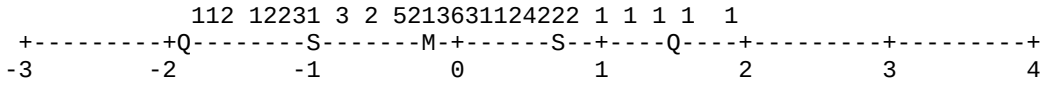
Infit MnSq:



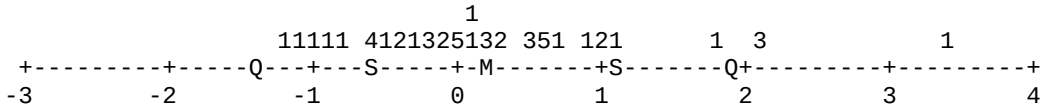
Outfit MnSq:



Infit ZStd:

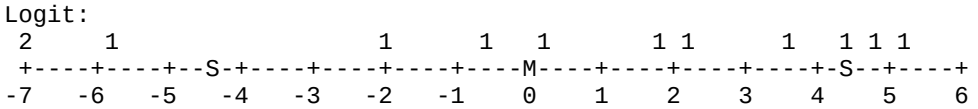


Outfit ZStd:

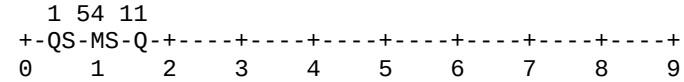


GMC 2012 Speaking Data 22/10/2012 21:37:47
Table 6.2 Testees Facet Summary.

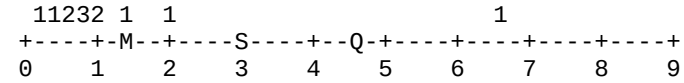
Logit:



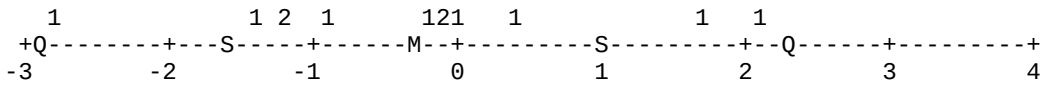
Infit MnSq:



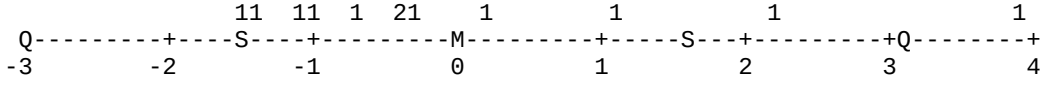
Outfit MnSq:



Infit ZStd:



Outfit ZStd:



GMC 2012 Speaking Data 22/10/2012 21:37:47
Table 7.1.1 Judges Measurement Report (arranged by N).

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Nu	Judges
32	12	2.67	2.85	-.75	.57	.76	-.3	.46	-.6	1.44	.94	.95	1	LP1
32	12	2.67	2.85	-.75	.57	.77	-.3	.47	-.6	1.37	.96	.95	2	LP2
31	12	2.58	2.69	-1.07	.58	.33	-1.6	.24	-1.0	1.57	.97	.95	3	LP3
33	12	2.75	3.01	-.43	.56	1.00	.1	.58	-.4	1.26	.94	.95	4	LP4
29	12	2.42	2.31	-1.74	.58	1.28	.6	.98	.3	.65	.94	.95	5	LP5
30	12	2.50	2.51	-1.41	.58	.45	-1.2	.31	-.7	1.62	.97	.95	6	LP6
36	12	3.00	3.40	.49	.54	.59	-.8	.41	-.6	1.50	.94	.93	7	NP1
34	12	2.83	3.15	-.12	.56	.91	.0	1.23	.5	.73	.95	.94	8	NP2
34	12	2.83	3.15	-.12	.56	1.25	.6	1.07	.3	.69	.93	.94	9	NP3
34	12	2.83	3.15	-.12	.56	1.61	1.1	.91	.1	.79	.92	.94	10	NP4
34	12	2.83	3.15	-.12	.56	1.35	.7	1.18	.4	.54	.95	.94	11	NP5
37	12	3.08	3.52	.78	.54	.87	-.1	.59	-.1	1.27	.93	.92	12	NP6
35	12	2.92	3.28	.19	.55	.49	-1.0	.86	.0	1.10	.97	.94	13	NP7
35	12	2.92	3.28	.19	.55	1.17	.4	.81	.0	1.04	.92	.94	14	WP1
35	12	2.92	3.28	.19	.55	.45	-1.2	.33	-.9	1.51	.95	.94	15	WP2
35	12	2.92	3.28	.19	.55	1.17	.4	.81	.0	1.04	.92	.94	16	WP3
34	12	2.83	3.15	-.12	.56	1.10	.3	.77	.0	1.03	.93	.94	17	WP4
36	12	3.00	3.40	.49	.54	1.19	.5	.81	.0	1.11	.91	.93	18	WP5
34	11	3.09	3.45	.61	.59	.85	-.1	.62	-.1	1.24	.93	.94	19	WP6
33	12	2.75	3.01	-.43	.56	1.92	1.5	1.80	1.1	.29	.92	.95	21	EAN1
29	12	2.42	2.31	-1.74	.58	.42	-1.3	.31	-.5	1.43	.97	.95	22	EAN2
33	12	2.75	3.01	-.43	.56	2.20	1.9	1.71	1.0	.39	.90	.95	23	EAN3
29	12	2.42	2.31	-1.74	.58	.84	-.1	.77	.1	.96	.94	.95	24	EAN4
28	12	2.33	2.11	-2.08	.59	.65	-.6	.67	.1	1.04	.95	.94	25	EAN5
46	12	3.83	4.27	3.18	.53	1.15	.4	.92	.9	.89	.79	.84	26	NIN1
46	12	3.83	4.27	3.18	.53	.76	-.4	.74	.8	1.11	.87	.84	27	NIN2
45	12	3.75	4.19	2.91	.52	.54	-1.0	.56	.6	1.29	.88	.85	28	NIN3
44	12	3.67	4.12	2.65	.51	.76	-.4	.65	.5	1.22	.85	.86	29	NIN4
44	12	3.67	4.12	2.65	.51	.76	-.4	.65	.5	1.22	.85	.86	30	NIN5
30	12	2.50	2.51	-1.41	.58	.75	-.3	.78	.0	1.10	.95	.95	31	LN1
29	12	2.42	2.31	-1.74	.58	.93	.0	.79	.1	.95	.95	.95	33	LN3
30	12	2.50	2.51	-1.41	.58	.90	.0	.85	.1	.87	.96	.95	35	LN5
38	12	3.17	3.62	1.06	.53	.61	-.7	.49	-.2	1.40	.92	.92	36	EAD1
39	12	3.25	3.72	1.34	.52	1.08	.3	.76	.1	1.00	.90	.91	37	EAD2
39	12	3.25	3.72	1.34	.52	.36	-1.6	.32	-.3	1.68	.92	.91	38	EAD3
37	12	3.08	3.52	.78	.54	1.37	.8	.92	.2	.96	.90	.92	39	EAD4
42	12	3.50	3.97	2.13	.51	.56	-1.0	.39	.1	1.48	.89	.88	40	EAD5
43	12	3.58	4.04	2.39	.51	1.10	.3	.84	.5	.86	.90	.87	41	LD1
44	12	3.67	4.12	2.65	.51	.98	.1	8.28	2.1	.17	.88	.86	42	LD2
43	12	3.58	4.04	2.39	.51	.92	.0	.79	.5	.95	.92	.87	43	LD3
44	12	3.67	4.12	2.65	.51	.98	.1	8.28	2.1	.17	.88	.86	44	LD4
40	12	3.33	3.81	1.61	.52	.38	-1.6	.34	-.2	1.34	.98	.90	45	LD5
35	12	2.92	3.28	.19	.55	1.32	.7	.89	.1	.96	.91	.94	47	SD2
38	12	3.17	3.62	1.06	.53	.92	.0	.69	.0	1.14	.91	.92	48	SD3
36	12	3.00	3.40	.49	.54	.97	.1	.88	.1	.90	.92	.93	49	SD4
38	12	3.17	3.62	1.06	.53	.62	-.7	.48	-.2	1.52	.92	.92	50	SD5
31	12	2.58	2.69	-1.07	.58	.75	-.3	.47	-.5	1.27	.96	.95	51	AH1
36	12	3.00	3.40	.49	.54	.87	.0	.74	.0	1.08	.93	.93	52	AH2
36	12	3.00	3.40	.49	.54	.40	-1.4	.27	-.9	1.53	.96	.93	53	AH3
35	12	2.92	3.28	.19	.55	1.22	.5	.86	.0	.89	.93	.94	54	AH4
35	12	2.92	3.28	.19	.55	.30	-1.7	.23	-1.2	1.64	.96	.94	55	AH5
44	12	3.67	4.12	2.65	.51	.49	-1.2	7.93	2.0	.61	.84	.86	57	RO2
42	12	3.50	3.97	2.13	.51	.54	-1.1	.48	.2	1.42	.90	.88	58	RO3
44	12	3.67	4.12	2.65	.51	.72	-.5	.70	.6	1.08	.90	.86	59	RO4
41	12	3.42	3.89	1.87	.51	1.43	.9	4.43	1.7	-.05	.82	.89	60	RO5
44	12	3.67	4.12	2.65	.51	.92	.0	.74	.6	1.12	.85	.86	61	RO6
40	12	3.33	3.81	1.61	.52	1.70	1.4	9.00	3.4	-1.01	.77	.90	62	RO7
36.7	12.0	3.06	3.38	.61	.55	.91	-.2	1.31	.2		.92		Mean (Count: 57)	
5.1	.1	.42	.58	1.46	.03	.39	.9	2.03	.8		.04		S.D. (Population)	
5.1	.1	.43	.59	1.48	.03	.40	.9	2.04	.8		.04		S.D. (Sample)	

Model, Populn: RMSE .55 Adj (True) S.D. 1.36 Separation 2.49 Strata 3.65 Reliability .86
Model, Sample: RMSE .55 Adj (True) S.D. 1.37 Separation 2.52 Strata 3.69 Reliability .86
Model, Fixed (all same) chi-square: 411.4 d.f.: 56 significance (probability): .00
Model, Random (normal) chi-square: 51.1 d.f.: 55 significance (probability): .63

GMC 2012 Speaking Data 22/10/2012 21:37:47
Table 7.2.1 Testees Measurement Report (arranged by N).

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Nu	Testees
204	57	3.58	3.64	-.50	.22	.78	-1.1	.75	-1.3	1.26	.76	.66	1	1
134	56	2.39	2.41	2.18	.20	.79	-1.2	.82	-1.0	1.25	.81	.73	2	2
282	57	4.95	4.98	-6.94	.61	.88	.0	.30	-.2	1.11	.34	.25	3	3
103	57	1.81	1.66	3.53	.21	.81	-.9	.89	-.3	1.23	.75	.70	4	4
233	57	4.09	4.11	-2.00	.24	1.35	1.6	1.46	2.1	.52	.52	.64	5	5
148	57	2.60	2.65	1.76	.19	.96	-.1	.93	-.3	1.05	.72	.73	6	6
86	57	1.51	1.33	4.40	.24	.71	-1.4	.64	-1.0	1.29	.73	.63	7	7
277	57	4.86	4.92	-5.73	.41	1.69	2.1	6.60	3.8	.13	-.08	.39	8	8
74	57	1.30	1.16	5.22	.29	1.08	.3	.63	-.6	1.13	.58	.54	9	9
187	57	3.28	3.35	.24	.20	.97	.0	1.02	.1	.98	.54	.68	10	10
80	57	1.40	1.24	4.78	.26	.47	-2.7	.45	-1.5	1.34	.75	.59	11	11
282	57	4.95	4.98	-6.94	.61	.91	.0	2.09	1.0	.98	.22	.25	12	12
174.2	56.9	3.06	3.04	.00	.31	.95	-.3	1.38	.0		.55		Mean (Count: 12)	
77.7	.3	1.36	1.44	4.31	.15	.30	1.3	1.64	1.5		.26		S.D. (Population)	
81.1	.3	1.42	1.50	4.51	.15	.31	1.3	1.71	1.6		.27		S.D. (Sample)	

Model, Populn: RMSE .34 Adj (True) S.D. 4.30 Separation 12.62 Strata 17.16 Reliability .99
Model, Sample: RMSE .34 Adj (True) S.D. 4.49 Separation 13.18 Strata 17.91 Reliability .99
Model, Fixed (all same) chi-square: 1601.1 d.f.: 11 significance (probability): .00
Model, Random (normal) chi-square: 10.9 d.f.: 10 significance (probability): .37

GMC 2012 Speaking Data 22/10/2012 21:37:47
Table 8.1 Category Statistics.

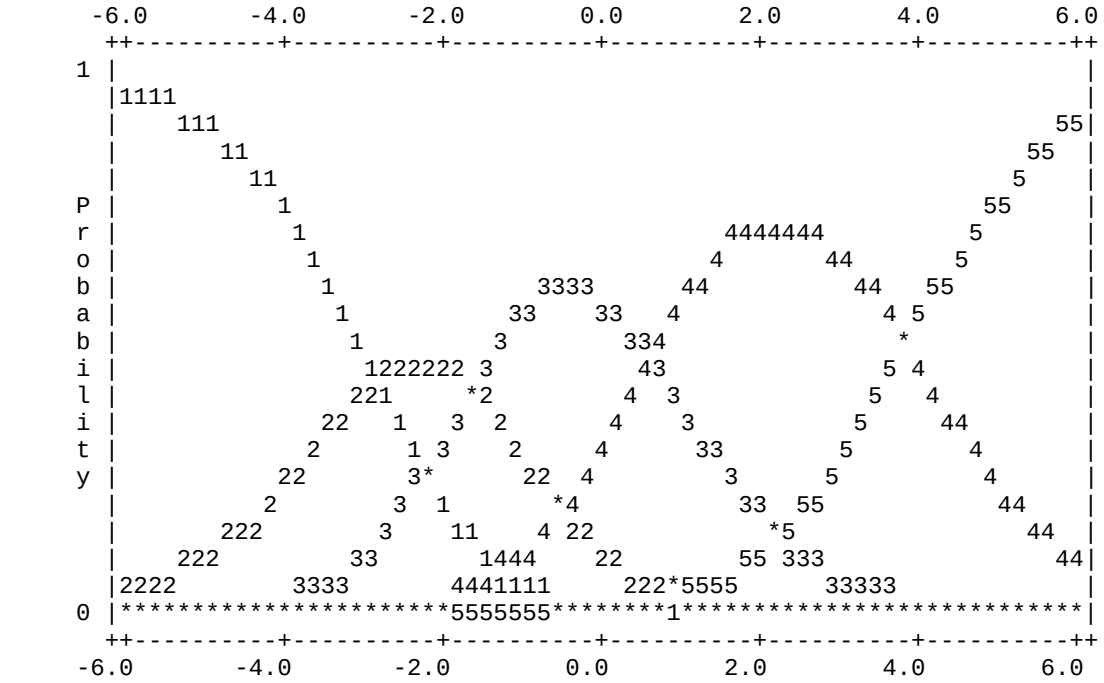
Model = ?,?,,R5

DATA Category Score	Counts Used	%	Cum. %	QUALITY CONTROL Avge Meas	Exp. Meas	OUTFIT MnSq	RASCH-ANDRICH Thresholds Measure	S.E.	EXPECTATION Measure at Category	-0.5	MOST PROBABLE from	RASCH-THURSTONE Thresholds	Cat PEAK Prob
1	169	25%	25%	-4.45	-4.34	.8			(-4.07)		low	low	100%
2	100	15%	39%	-2.24	-2.39	.6	-2.82	.15	-2.28	-3.27	-2.82	-3.04	47%
3	117	17%	57%	-.53	-.54	1.1	-1.63	.15	-.43	-1.41	-1.63	-1.51	59%
4	115	17%	73%	2.11	1.96	2.4	.60	.16	2.24	.73	.60	.65	72%
5	182	27%	100%	6.68	6.76	1.0	3.85	.20	(4.96)	3.94	3.85	3.88	100%
									(Mean)		(Modal)	(Median)	

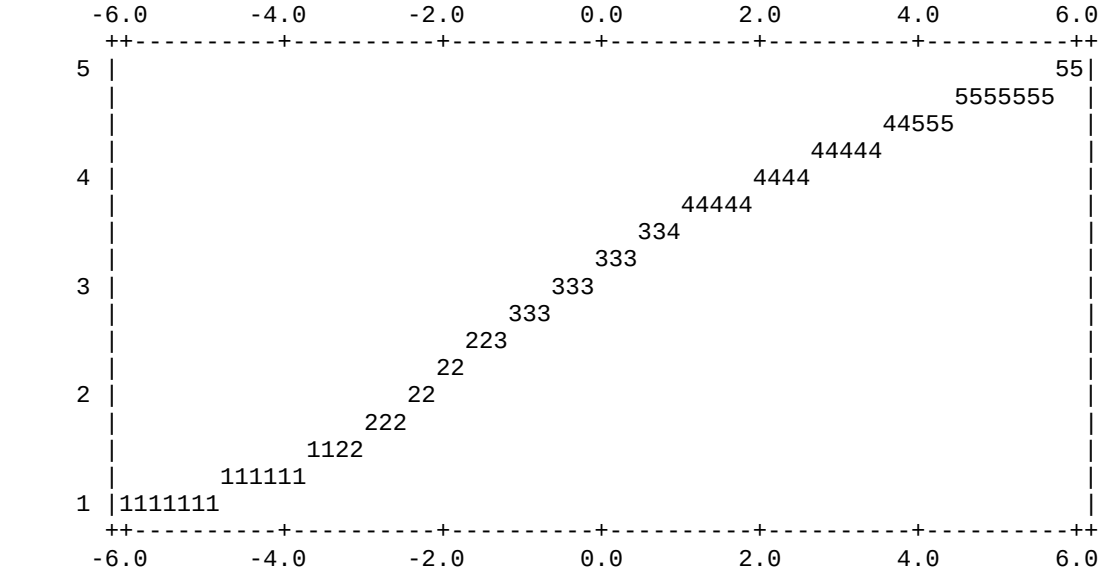
Scale structure

```
Measr:-6.0       -4.0       -2.0        0.0        2.0        4.0        6.0
        +          +          +          +          +          +          +
  Mode:<1---------(^)---12--^--23------^---34--------^-------45----(^)----5>

Median:<1---------(^)--12---^---23-----^----34-------^-------45----(^)----5>

  Mean:<1---------(^)-12----^---23-----^----34-------^--------45---(^)----5>
        +          +          +          +          +          +          +
Measr:-6.0       -4.0       -2.0        0.0        2.0        4.0        6.0
```

Probability Curves



Expected Score Ogive (Model ICC)



GMC 2012 Speaking Data 22/10/2012 21:37:47
Table 4.1 Unexpected Responses (8 residuals sorted by u).

Cat	Score	Exp.	Resd	StRes	Nu	Judg	Nu	Te
4	4	5.0	-1.0	-9.0	42	LD2	8	8
4	4	5.0	-1.0	-9.0	44	LD4	8	8
4	4	5.0	-1.0	-9.0	57	RO2	8	8
4	4	5.0	-1.0	-9.0	62	RO7	12	12
4	4	5.0	-1.0	-6.5	60	RO5	8	8
4	4	5.0	-1.0	-5.7	62	RO7	8	8
3	3	1.3	1.7	3.4	21	EAN1	4	4
3	3	1.3	1.7	3.4	23	EAN3	4	4
Cat	Score	Exp.	Resd	StRes	Nu	Judg	Nu	Te

Appendix 5b: Many Facet Rasch Output – Writing

```
Facets (Many-Facet Rasch Measurement) Version No. 3.70.1  Copyright(c) 1987-
2012, John M. Linacre. All rights reserved.
07/11/2012 19:50:41
GMC 2012 Writing Data 07/11/2012 19:50:41
Table 1. Specifications from file "G:\GMC Project\gmcwritedat4.txt".

Title = GMC 2012 Writing Data 07/11/2012 19:50:41
Data file = (G:\GMC Project\gmcwritedat4.txt)
Output file = gmcwrite4.out

; Data specification
Facets = 3
Delements = N
Non-centered =  1
Positive =  1
Labels =
 1,Judges ; (elements = 54)
 2,Candidates ; (elements = 8)
 3,Group ; (elements = 5)
Model = ?,?,,R5,1

; Output description
Arrange tables in order = N
Bias/Interaction direction = plus ; ability, easiness, leniency: higher score =
positive logit
Fair score = Mean
Pt-biserial = Measure
Heading lines in output data files = Y
Omit unobserved elements = yes
Barchart = Yes
Total score for elements = Yes
T3onscreen show only one line on screen iteration report = Y
T4MAX maximum number of unexpected observations reported in Table 4 = 100
T8NBC show table 8 numbers-barcharts-curves = NBC
Unexpected observations reported if standardized residual >= 3
Usort unexpected observations sort order = u
WHexact - Wilson-Hilferty standardization = Y
; Convergence control
Convergence = .5, .01
Iterations (maximum) = 0 ; unlimited
Xtreme scores adjusted by = .3, .5  ;(estimation, bias)

GMC 2012 Writing Data 07/11/2012 19:50:41
Table 2. Data Summary Report.

Assigning models to Data= "G:\GMC Project\gmcwritedat4.txt"

In data: 1, 1, 1, 1
Check (1)? Unspecified element: facet: 1 element: 1 1 in line 84 - see Table 2.
Delements = N

Total lines in data file = 498
Total data lines = 496
Responses matched to model: ?,?,,R5,1 = 432
    Total non-blank responses found = 496
Responses with unspecified elements = 64
Number of blank lines = 2
Valid responses used for estimation = 432
```

List of unspecified elements. Please copy and paste into your specification file, where needed
Labels=Nobuild ; to suppress this list 1, Judges, ; facet 1
 1 = 1
 25 = 25
 26 = 26
 36 = 36
 37 = 37
 49 = 49
 61 = 61
 9 = 9
*

GMC 2012 Writing Data 07/11/2012 19:50:41
Table 3. Iteration Report.

Iteration		Max. Score Residual Elements	%	Categories	Max. Logit Change Elements	Steps
PROX	1				-2.5683	
PROX	2				.3405	
JMLE	3	19.5566	9.1	-79.5990	.4929	1.5872
JMLE	4	-10.1496	-4.7	-11.4050	-.2886	.1173
JMLE	5	6.3884	3.0	-7.8702	-.2055	.1125
JMLE	6	4.9545	2.3	-5.6093	-.1588	.0931
JMLE	21	.9362	.6	.8907	-.0259	.0198
*						
JMLE	28	.5248	.3	.4992	-.0152	.0118
JMLE	29	.4831	.3	.4595	-.0141	.0110
JMLE	30	.4447	.3	.4230	-.0131	.0102
JMLE	31	.4092	.2	.3892	-.0122	.0095
JMLE	32	.3764	.2	.3580	-.0113	.0088
JMLE	33	.3462	.2	.3292	-.0105	.0082
JMLE	34	.3183	.2	.3026	-.0097	.0076

Subset connection O.K.

GMC 2012 Writing Data 07/11/2012 19:50:41

Table 4. Unexpected Responses - appears after Table 8.

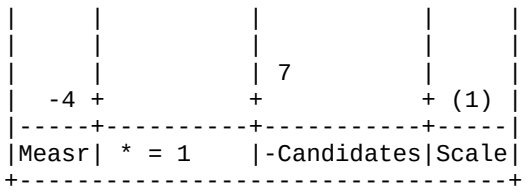
GMC 2012 Writing Data 07/11/2012 19:50:41
Table 5. Measurable Data Summary.

Cat	Score	Exp.	Resd	StRes	
3.04	3.04	3.04	.00	.01	Mean (Count: 432)
1.35	1.35	1.14	.71	.96	S.D. (Population)
1.35	1.35	1.14	.71	.96	S.D. (Sample)

Data log-likelihood chi-square = 841.7791

Approximate degrees of freedom = 369

```
Chi-square significance prob.  = .0000
                                          Count   Mean   S.D.   Params
Responses used for estimation      =        432   3.04   1.35       63
Count of measurable responses      =        432
Raw-score variance of observations   =   1.83 100.00%
Variance explained by Rasch measures =   1.33  72.71%
Variance of residuals                =   0.50  27.29%


GMC 2012 Writing Data 07/11/2012 19:50:41
Table 6.0  All Facet Vertical "Rulers".

Vertical = (1*,2A,3A,S) Yardstick (columns lines low high extreme)= 0,7,-
4,3,End
+-----------------------------------+
|Measr|+Judges   |-Candidates|Scale|
|-----+----------+-----------+-----|
|   3 +          +           + (5) |
|     |          |           | --- |
|     |          |           |     |
|     |          |           |     |
|     |          | 1         |     |
|     |          |           |     |
|     | *        |           |     |
|   2 +          + 2         +     |
|     | ***      |           |     |
|     |          |           |     |
|     | ***      |           |     |
|     |          | 4         |  4  |
|     | *        |           |     |
|     | *        |           |     |
|   1 +          +           +     |
|     |          |           |     |
|     |          |           |     |
|     | ******   |           |     |
|     | **       |           |     |
|     |          |           | --- |
|     | *****    | 5         |     |
*   0 * ****     *           *     *
|     |          | 3         |     |
|     | *******  |           |     |
|     | ******** |           |  3  |
|     |          |           |     |
|     | ******   |           |     |
|     |          | 6         |     |
|  -1 + *****    +           + --- |
|     | *        | 8         |     |
|     |          |           |     |
|     |          |           |     |
|     |          |           |  2  |
|     |          |           |     |
|     | *        |           |     |
|  -2 +          +           +     |
|     |          |           |     |
|     |          |           | --- |
|     |          |           |     |
|     |          |           |     |
|     |          |           |     |
|     |          |           |     |
|  -3 +          +           +     |
|     |          |           |     |
|     |          |           |     |
|     |          |           |     |
```



S.1: Model = ?,?,,R5

GMC 2012 Writing Data 07/11/2012 19:50:41

Table 6.1 Judges Facet Summary.

Logit:

```
                     1      1  5 6 8 7  4 5 2 6    1 1  3  3  1
 +---------+---------+--Q------+-S-------+M-------S+-------Q-+---------+
-4        -3        -2        -1         0         1         2         3
```

Infit MnSq:

```
    1214 31 241 11141422152  1311111 11
 Q--------S---------M+-------S--------Q--+-------------------+
 0                   1                   2                   3
```

Outfit MnSq:

```
    13 312232  12 21 65 233323 1  1     1    1
 Q--------S---------M+-------S--------Q--+-------------------+
 0                   1                   2                   3
```

Infit ZStd:

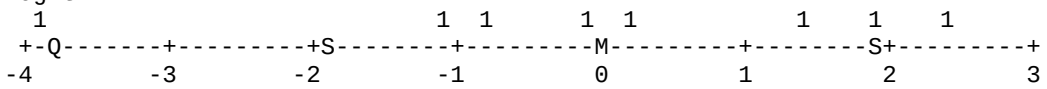
```
      11   2 12 112  1 241 1 213142334  412 2 2
 +-------Q-+-------S-+--------M+--------S+--------Q+---------+
-3        -2        -1         0         1         2         3
```

Outfit ZStd:

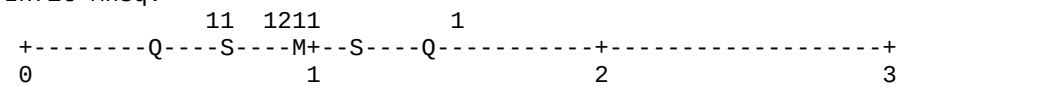
```
          111 1 1 311321   13 21 83125621  1  1  1
 +---------+Q--------S--------M+-------S-+------Q--+---------+
-3        -2        -1         0         1         2         3
```

GMC 2012 Writing Data 07/11/2012 19:50:41
Table 6.2 Candidates Facet Summary.

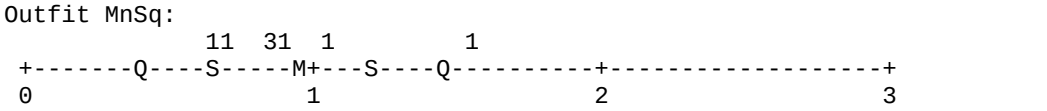
Logit:



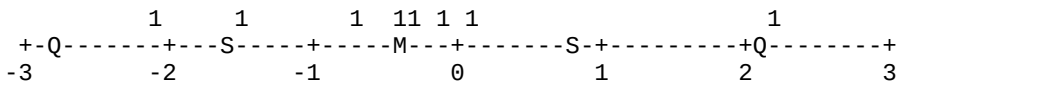
Infit MnSq:



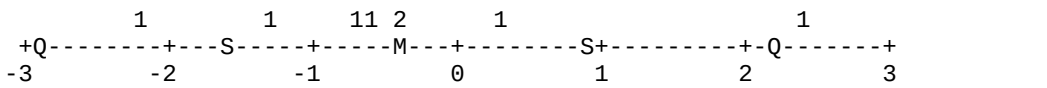
Outfit MnSq:



Infit ZStd:



Outfit ZStd:



GMC 2012 Writing Data 07/11/2012 19:50:41
Table 7.1.1 Judges Measurement Report (arranged by N).

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Nu	Judges
20	8	2.50	2.54	-.93	.48	1.15	.4	1.05	.2	.63	.83	.86	2	LP2
16	8	2.00	1.73	-1.88	.51	1.55	1.0	1.42	.7	.64	.77	.84	3	LP3
20	8	2.50	2.54	-.93	.48	.28	-1.7	.31	-1.2	1.74	.96	.86	4	LP4
31	8	3.88	4.06	1.58	.51	.22	-2.0	.22	-1.5	1.84	.91	.73	5	LP5
27	8	3.38	3.66	.64	.47	.76	-.3	.75	-.3	1.15	.82	.80	6	LP6
27	8	3.38	3.66	.64	.47	.20	-2.3	.22	-1.9	1.61	.92	.80	7	NP1
21	8	2.63	2.75	-.71	.48	1.07	.3	.88	.0	.90	.85	.85	8	NP2
23	8	2.88	3.12	-.25	.48	1.07	.3	1.23	.5	.67	.74	.84	10	NP4
22	8	2.75	2.94	-.48	.48	1.12	.3	1.16	.4	.80	.76	.85	11	NP5
26	8	3.25	3.54	.41	.47	.32	-1.7	.30	-1.6	1.73	.90	.81	12	NP6
27	8	3.38	3.66	.64	.47	.66	-.6	.49	-.9	1.54	.82	.80	13	NP7
20	8	2.50	2.54	-.93	.48	.80	-.2	1.03	.2	.84	.79	.86	14	WP1
22	8	2.75	2.94	-.48	.48	.96	.1	1.06	.2	1.07	.85	.85	15	WP2
25	8	3.13	3.41	.19	.47	1.19	.5	1.26	.6	.58	.80	.82	16	WP3
22	8	2.75	2.94	-.48	.48	1.52	.9	1.22	.5	.74	.78	.85	17	WP4
21	8	2.63	2.75	-.71	.48	1.42	.8	1.26	.5	.64	.80	.85	18	WP5
21	8	2.63	2.75	-.71	.48	.89	.0	.99	.1	.61	.79	.85	19	WP6
21	8	2.63	2.75	-.71	.48	1.20	.5	1.35	.7	.27	.73	.85	20	WP7
25	8	3.13	3.41	.19	.47	.88	.0	1.22	.5	.98	.74	.82	21	EAN1
19	8	2.38	2.33	-1.16	.48	.31	-1.6	.31	-1.1	1.65	.96	.86	22	EAN2
24	8	3.00	3.27	-.03	.47	1.21	.5	1.29	.6	.53	.71	.83	23	EAN3
20	8	2.50	2.54	-.93	.48	1.01	.2	1.32	.6	.28	.76	.86	24	EAN4
33	8	4.13	4.28	2.16	.57	.46	-.9	.53	-.4	1.42	.77	.70	27	NIN2
29	8	3.63	3.86	1.09	.48	.92	.0	1.02	.2	.66	.79	.77	28	NIN3
32	8	4.00	4.17	1.86	.54	1.23	.5	1.13	.4	.78	.66	.72	29	NIN4
31	8	3.88	4.06	1.58	.51	1.23	.5	.98	.1	1.09	.72	.73	30	NIN5
27	8	3.38	3.66	.64	.47	.26	-2.0	.22	-1.9	1.92	.89	.80	31	LN1
23	8	2.88	3.12	-.25	.48	1.46	.9	1.19	.5	.84	.80	.84	32	LN2
22	8	2.75	2.94	-.48	.48	.32	-1.5	.36	-1.3	1.55	.94	.85	33	LN3
26	8	3.25	3.54	.41	.47	1.67	1.2	1.39	.8	.63	.68	.81	34	LN4
27	8	3.38	3.66	.64	.47	1.80	1.4	2.21	1.8	-.17	.50	.80	35	LN5
25	8	3.13	3.41	.19	.47	1.11	.3	1.03	.2	.99	.80	.82	38	EAD3
23	8	2.88	3.12	-.25	.48	1.43	.8	1.94	1.5	.34	.67	.84	39	EAD4
23	8	2.88	3.12	-.25	.48	.99	.1	1.29	.6	1.09	.82	.84	40	EAD5
32	8	4.00	4.17	1.86	.54	1.01	.2	.98	.2	.94	.61	.72	41	LD1
31	8	3.88	4.06	1.58	.51	.59	-.6	.41	-.9	1.55	.80	.73	42	LD2
30	8	3.75	3.96	1.33	.50	.15	-2.5	.14	-2.0	1.92	.97	.75	43	LD3
32	8	4.00	4.17	1.86	.54	1.01	.2	.98	.2	.94	.61	.72	44	LD4
24	8	3.00	3.27	-.03	.47	.59	-.7	.87	.0	.71	.79	.83	45	LD5
21	8	2.63	2.75	-.71	.48	.55	-.8	.72	-.3	1.19	.90	.85	46	SD1
22	8	2.75	2.94	-.48	.48	.58	-.7	.55	-.7	1.63	.90	.85	47	SD2
22	8	2.75	2.94	-.48	.48	1.44	.8	1.39	.7	.16	.67	.85	48	SD3
20	8	2.50	2.54	-.93	.48	1.70	1.2	1.49	.8	-.12	.74	.86	50	SD5
22	8	2.75	2.94	-.48	.48	.40	-1.2	.44	-1.0	1.47	.92	.85	51	AH1
24	8	3.00	3.27	-.03	.47	.38	-1.3	.39	-1.2	1.40	.91	.83	52	AH2
23	8	2.88	3.12	-.25	.48	.54	-.8	.48	-.9	1.51	.91	.84	53	AH3
25	8	3.13	3.41	.19	.47	.61	-.7	.48	-1.0	1.68	.87	.82	54	AH4
23	8	2.88	3.12	-.25	.48	.83	-.1	.74	-.3	1.50	.86	.84	55	AH5
25	8	3.13	3.41	.19	.47	1.22	.5	1.03	.2	1.12	.83	.82	56	RO1
22	8	2.75	2.94	-.48	.48	1.60	1.1	1.33	.7	.90	.70	.85	57	RO2
23	8	2.88	3.12	-.25	.48	1.87	1.4	1.65	1.1	.59	.79	.84	58	RO3
21	8	2.63	2.75	-.71	.48	.38	-1.3	.46	-.9	1.71	.87	.85	59	RO4
24	8	3.00	3.27	-.03	.47	.89	.0	.85	-.1	1.25	.82	.83	60	RO5
27	8	3.38	3.66	.64	.47	1.20	.5	1.01	.1	1.17	.79	.80	62	RO7
24.3	8.0	3.04	3.23	.05	.48	.93	-.1	.93	-.1		.80		Mean (Count: 54)	
3.9	.0	.48	.54	.89	.02	.46	1.0	.46	.9		.10		S.D. (Population)	
3.9	.0	.49	.55	.90	.02	.46	1.0	.46	.9		.10		S.D. (Sample)	

Model, Populn: RMSE .48 Adj (True) S.D. .75 Separation 1.54 Strata 2.39 Reliability .70
Model, Sample: RMSE .48 Adj (True) S.D. .76 Separation 1.56 Strata 2.41 Reliability .71
Model, Fixed (all same) chi-square: 165.7 d.f.: 53 significance (probability): .00
Model, Random (normal) chi-square: 42.3 d.f.: 52 significance (probability): .83

GMC 2012 Writing Data 07/11/2012 19:50:41
Table 7.2.1 Candidates Measurement Report (arranged by N).

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	N	Candidates
87	54	1.61	1.47	2.36	.20	1.00	.0	1.05	.2	1.00	.59	.61	1	1
99	54	1.83	1.72	1.94	.18	.71	-1.5	.71	-1.2	1.13	.69	.63	2	2
178	54	3.30	3.39	-.09	.17	1.48	2.2	1.54	2.4	.40	.50	.54	3	3
117	54	2.17	2.13	1.43	.16	.86	-.7	.91	-.4	1.19	.66	.63	4	4
169	54	3.13	3.23	.15	.16	.65	-2.1	.63	-2.1	1.44	.62	.56	5	5
200	54	3.70	3.76	-.79	.19	.89	-.4	.85	-.6	1.05	.66	.50	6	6
256	54	4.74	4.78	-3.91	.31	.97	.0	.85	-.4	1.04	.28	.31	7	7
208	54	3.85	3.89	-1.10	.20	.92	-.3	.86	-.6	1.04	.23	.49	8	8
164.3	54.0	3.04	3.05	.00	.20	.94	-.4	.93	-.4		.53		Mean (Count: 8)	
55.1	.0	1.02	1.09	1.89	.05	.24	1.2	.26	1.3		.17		S.D. (Population)	
58.9	.0	1.09	1.16	2.02	.05	.25	1.3	.28	1.3		.18		S.D. (Sample)	

Model, Populn: RMSE .20 Adj (True) S.D. 1.88 Separation 9.31 Strata 12.75 Reliability .99
Model, Sample: RMSE .20 Adj (True) S.D. 2.01 Separation 9.96 Strata 13.62 Reliability .99Model, Fixed (all same)
chi-square: 511.3 d.f.: 7 significance (probability): .00
Model, Random (normal) chi-square: 6.9 d.f.: 6 significance (probability): .33

GMC 2012 Writing Data 07/11/2012 19:50:41
Table 8.1 Category Statistics.

Model = ?,?,,R5

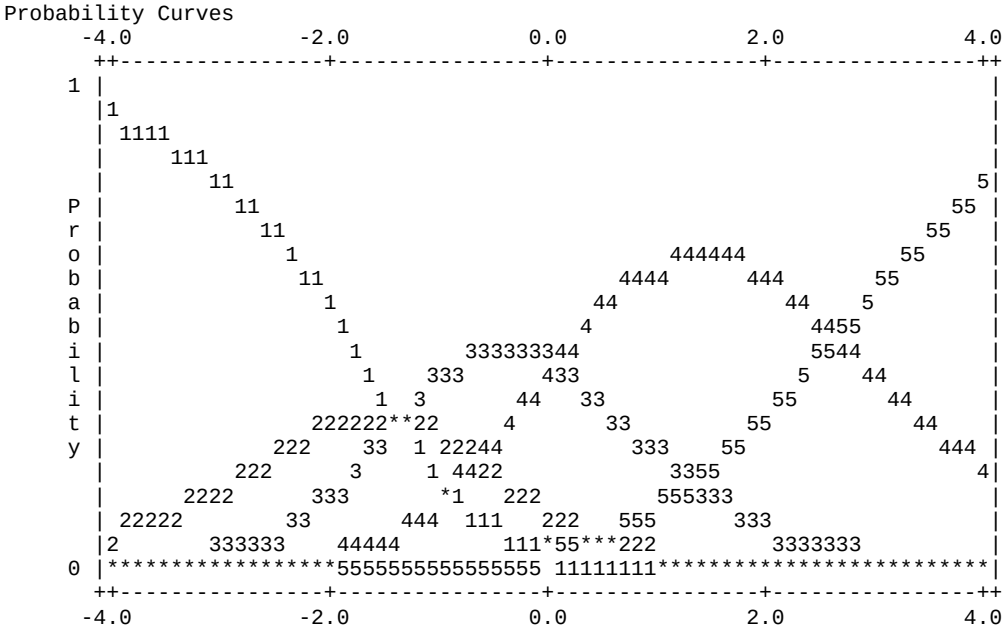
DATA Category Score	Counts Used	%	Cum. %	QUALITY CONTROL Avge Meas	Exp. Meas	OUTFIT MnSq	RASCH-ANDRICH Thresholds Measure	S.E.	EXPECTATION Measure at Category	-0.5	MOST PROBABLE from	RASCH-THURSTONE Thresholds	Cat PEAK Prob
1	88	20%	20%	-2.21	-2.21	.9			(-2.88)		low	low	100%
2	58	13%	34%	-1.48	-1.34	1.0	-1.38	.17	-1.53	-2.25	-1.38	-1.89	32%
3	98	23%	56%	-.29	-.28	.8	-1.33	.16	-.40	-.98	-1.33	-1.07	45%
4	124	29%	85%	1.10	.95	.9	.05	.14	1.42	.35	.05	.19	64%
5	64	15%	100%	3.04	3.19	1.2	2.66	.19	(3.79)	2.83	2.66	2.72	100%

(Mean) (Modal) (Median)

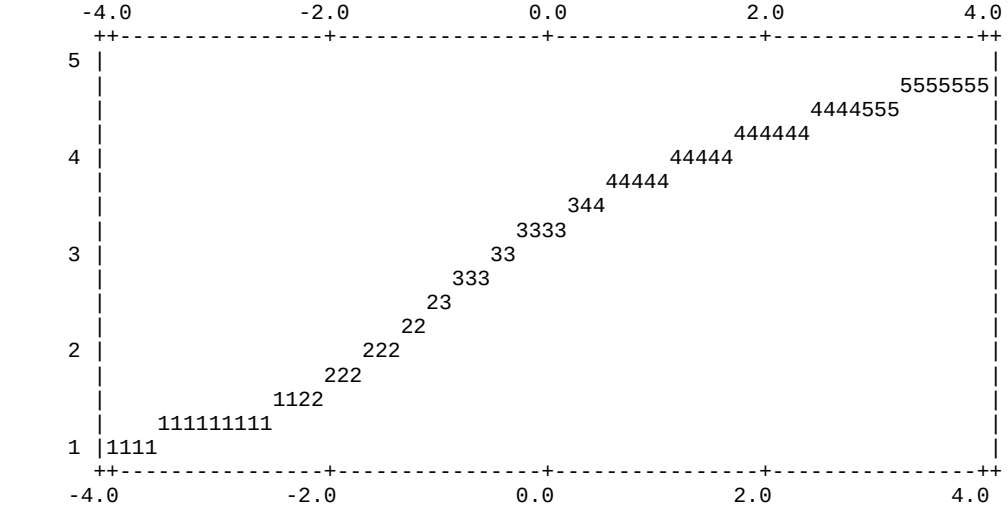
Scale structure

```
Measr:-4.0             -2.0              0.0              2.0              4.0
        +                +                +                +                +
  Mode:<1---------(^)--------123-------^-34-----------^---------45-------(^)5>

Median:<1---------(^)----12--^--23----^----34---------^---------45-------(^)5>

  Mean:<1--------(^)--12-----^---23----^----34--------^----------45------(^)5>
        +                +                +                +                +
Measr:-4.0             -2.0              0.0              2.0              4.0
```

Probability Curves



Expected Score Ogive (Model ICC)



GMC 2012 Writing Data 07/11/2012 19:50:41
Table 4.1 Unexpected Responses (0 residuals sorted by u).

Cat	Score	Exp.	Resd	StRes	Nu	Judg	N	C
*** No unexpected observation with StRes >= 3								

Appendix 6a: IELTS requirements for UK professionals

Dentists

http://www.rcseng.ac.uk/fds/nacpde/overseas_qualified/english_language.html

The GDC requirements before sitting the Overseas Registration Exam (ORE) are as follows
An academic IELTS test reports which must:

- Be no more than two years old on the date of receipt at the GDC
- Show a minimum overall band score of 7.0
- Score no lower than 6.5 in any section

Vets

http://www.rcvs.org.uk/registration/statutory-membership-exam/

RCVS English language competence requirements
Before applying to sit the statutory membership examination, all examination applicants must meet the RCVS requirement for proof of English language competence (see 'Related documents' - top right) by providing evidence of an overall band score of at least 7.0 in the academic International English Language Testing System (IELTS) English test.

Nurses and Midwives

http://www.nmc-uk.org/Registration/Joining-the-register/Trained-outside-the-EU--EEA/International-English-Language-Testing-IELTS/

All non EU trained applicants to the nurses or midwives part of the register must complete and provide evidence of the International English Language Test (IELTS) before submitting their application to the NMC. You must complete the academic version of the IELTS test and achieve:

- At least 7.0 in the listening and reading sections
- At least 7.0 in the writing and speaking sections
- At least 7.0 (out of a possible 9) overall

Optometrists

http://www.optical.org/en/our_work/Education/Criteria_for_registration/Optometrists/Non_EU_EEA_route_to_registration.cfm

The non EU/EEA route to registration as an optometrist in the UK

To qualify as an optometrist via this route you must:
1. Have undertaken three years' full time optometric training, either post baccalaureate or from the age of 18, and completed that qualification in the country of training.
2. Be legally qualified to practise optometry in a country outside of the UK.
3. Have completed a minimum of one year's unsupervised practice, post-qualification.
4. Obtain a score a minimum score of 7 in the International English Language Testing System (IELTS). Individual scores for each section of the test must not be lower than 6, with the exception of the 'Speaking' section, where you must score a minimum of 7.

5. Pass the non EU/ EEA examination run by the College of Optometrists. More information is available on their website.

Osteopaths

http://www.osteopathy.org.uk/practice/How-to-register-with-the-GOsC/Qualified-outside-the-UK/Applicants-from-outside-the-EUEEA-and-Switzerland/

The ability to communicate effectively with patients is critical to working as an osteopath in the United Kingdom (UK). The national language of the UK is English and the General Osteopathic Council (GOsC) requires registrants to be proficient in this language. As part of the application for registration, the GOsC requires all applicants from outside the European Union* to provide evidence of their English language ability. The preferred testing system of the GOsC is the IELTS system, at which applicants need to score 7.0 or above. Please note that we require you to take the IELTS 'Academic' test and not the 'General' test.

Allied Health Professionals

http://www.hpc-uk.org/apply/international/requirements/

English proficiency

Applicants whose first language is not English and who are required to provide a language test certificate as evidence of their proficiency must ensure that it is, or is comparable to, IELTS level 7.0 with no element below 6.5 (or IELTS level 8.0 with no element below 7.5 for Speech and language therapists).

* Speech and language therapists: this Standard applies to both EEA and International applicants. This requirement is higher for speech and language therapists than for all other professions, as communication in English is a core professional skill (see 1b.3 of the standards of proficiency).

Pharmacists/pharmacy technicians

http://www.pharmacyregulation.org/sites/default/files/International%20Information%20Pack%20Mar%202012.pdf

1. English Language Test
You MUST supply a satisfactory English language test result with your initial application. The General Pharmaceutical Council (GPhC) only accepts IELTS and requires a level of a minimum of 7 in every category of the same sitting of the academic IELTS . The IELTS result is valid for 2 years from the date of the test. Your IELTS result must be valid until your application is complete and considered for eligibility.

Bar Standards Board

http://www.barstandardsboard.org.uk/

Qualifying as a barrister – IELTS 7.5 with no individual score lower.

Appendix 6b: English language requirements for medical registration – overseas English speaking countries

Australia	IELTS minimum 7 in each of the components; or Grade A or B in each of the four subtests of the Occupational English Test (OET); or A pass in the Professional Linguistic Assessment Board (PLAB) examination in the United Kingdom; or A pass in the English language proficiency component of the New Zealand Registration Examination (NZREX); or A pass in the English language component of the United States Medical Licensing Examination (USMLE)
Canada	There is no general rule and different regions are required to set their own scores. The general minimum requirements are: IELTS minimum 7 in each of the components; or TOEFL academic iBT minimum of 24 in each component
Ireland	Overall IELTS Band 7, no score lower than 6.5
New Zealand	Overall IELTS 7.5 with minimum 7.5 for Speaking and Listening; minimum 7.0 for Writing and Reading
South Africa	Occupational English Test (OET) – grade A or B; or IELTS 7 or higher in each of the 4 components
United States of America	No language proficiency test score is required as all potential registrants (USA and Canada graduates and IMGs) are first required to pass Steps 1 and 2 of USMLE (the United States Medical Licensing Examination) after which they must obtain a residency position and then complete Step 3 of USMLE, approximately one year later. Language proficiency is assessed as part of the USMLE.