PLYMOUTH UNIVERSITY PENINSULA
SCHOOLS OF MEDICINE & DENTISTRY

A Systematic Review on the impact of licensing examinations for doctors in countries comparable to the UK

Final Report

Dr Julian Archer, Dr Nick Lynn, Mr Martin Roberts, Dr Lee Coombes, Dr Tom Gale and Dr Sam Regan de Bere

29/05/2015

CAMERA
Collaboration for the Advancement of Medical Education Research & Assessment

Funded by The General Medical Council.

The views expressed in this report are those of the authors and do not necessarily reflect those of the General Medical Council.

Table of Contents

Table of Figures

Table of Abbreviations and Acronyms

AERA	American Educational Research Association
APA	American Psychological Association
CAMERA	Collaboration for the Advancement of Medical Education Research
CEO	Chief Executive Officer
ECFMG	Educational Commission for Foreign Medical Graduates
EEA	European Economic Area
EU	European Union
FSMB	Federal State Medical Board
FAIMER	Foundation for Advancement of International Medical Education Research
GMC	General Medical Council
IMG	International Medical Graduate
NBME	National Board on Medical Examiners
PLAB	Professional and Linguistic Assessment Board
UNDP	United Nations Development Programme
USMLE	United States Medical Licensing Examination

Executive Summary

The General Medical Council (GMC), among its many regulatory functions, has a mandate to maintain the health and safety of the public in the United Kingdom. It does this by regulating doctors at all stages of their medical education and training, and their subsequent professional development. The GMC exists to ensure all doctors licensed to practise in the United Kingdom (UK) are of the highest standard.

In an increasingly globalized world physician shortages, economic incentives, and ease of travel are encouraging the physician workforce to greater mobility and increased migration. This makes the task of regulation more complex, in that it can be difficult to establish the quality and competence of doctors trained outside the GMC's jurisdiction. One way in which the GMC might be able to fully achieve its mandate to maintain the health and safety of the public is through the use of licensing examinations.

Licensing examinations, it is suggested, would provide a minimum standard of performance for doctors working in the UK and that they would introduce some standardisation into medical education and medical practice. National licensing examinations are also said to be useful in predicting the future performance of medical graduates. Together, greater patient safety might be achieved.

In an effort to assess the body of evidence that exists to support these claims and establish the validity of licensure examinations, the GMC commissioned the Collaboration for the Advancement of Medical Education Research and Assessment (CAMERA), to carry out a systematic review of the literature on licensing (or similar) examinations in the 49 countries currently seen as comparable to the UK (UNDP, 2014).

The review had three aims:

1. To establish the existing evidence base for the validity of medical licensing examinations or similar in countries comparable to the UK.
2. To establish the validity evidence for the impact of medical licensing examinations.
3. To identify best practice and any gaps in knowledge for medical licensing examinations.

Our search strategy involved interrogating seven international electronic databases. We used clearly defined inclusion and exclusion criteria that sought to capture academic research and grey literature published in any language since 2005. In addition we also searched the websites of the medical regulators or licensing bodies in each of the 49 countries (where available), and the websites of specialist assessment organisations.

Finally, in collaboration with the GMC and the International Association of Medical Regulatory Authorities (IAMRA), we surveyed world regulators and asked them to provide us with details of any research that informed their thinking. We did this in an effort to ensure we captured all relevant literature.

Whilst the review adhered to the protocols and procedures currently recognised as 'best practice' by systematic reviewers, a unique feature of this review, which was intended to ensure we met its specific aims, was the use of the American Psychological Association's validity framework (Downing, 2003).

The findings

From a total of 202 retrieved papers, only 73 fulfilled the initial criteria for the main review. However when these were mapped against the validity framework only 23 of the 73 offered any *empirical* evidence for the validity of licensure examinations. Many were concerned with the technical aspects of licensure examinations - the 'science' of assessment (Ahn & Ahn, 2007; Lillis, 2012; Ranney, 2006) - and it is here that the strongest validity evidence was found.

A number of papers demonstrated how performance in licensing examinations, primarily the United States Medical Licensing Examination (USMLE), features in later examination selection processes, and therefore has career consequences for doctors (Green, Jones & Thomas, 2009; Kenny, McInnes & Singh, 2013). Others highlighted performance differences between different groups of doctors. These show clearly that International Medical Graduates (IMGs) perform less well in national licensing examinations (Hecker & Violato, 2008; Margolis et al.; McManus & Wakeford, 2014). The research however does not establish why this might be or how these differences can be explained. Again, this has consequences for future careers (Musoke, 2012; Sonderen et al., 2009). One further study, which approached licensing examinations and physician migration from an economic perspective points to how, for a variety of economic and professional reasons, licensing examinations might dissuade some skilled practitioners from remaining in their profession (Kugler A.D & Sauer, 2005).

Some authors claim to provide evidence that licensing examinations ensure greater patient safety and improved quality of care (McMahon & Tallia, 2010; Melnick, 2009; Norcini et al., 2014; Tamblyn et al., 2007; Wenghofer et al., 2009). The evidence for these claims however is based on correlations of performance that fail to establish a direct link between national licensing examinations and improvements in patient outcomes. Notwithstanding the lack of a substantive causal link, these correlations are important in demonstrating the value of knowledge acquisition and the broader role for testing in medical education.

The remaining 50 papers, when mapped to the validity framework, revealed a general lack of *validity* evidence. They were, nevertheless, important. Most consisted of informed reasoning or opinion and were written by acknowledged experts in the field of educational and medical assessment. And while all drew on research material to argue their case, the evidence used was mostly equivocal. Indeed, some literature was used to argue both viewpoints. In the absence of any *compelling* validity evidence for licensing examinations the arguments for and against will continue unabated.

Conclusions

In short we conclude, the debate around licensure examinations is strong on opinion but weak on evidence. This is especially true of any wider claims that licensure examinations improve patient safety and practitioner competence (Sutherland, 2006).

What is clear from the literature is that where national licensing, like other large scale, examinations exist there is a correlation between a doctor's performance in the national licensing examination and their subsequent examination performances. There is also a correlation between licensing examinations and some patient outcomes and rates of complaints. However the question of whether the introduction of a national licensing examination would raise standards of medical practice is not addressed by these correlations. Introducing an end-of-medical-school national licensing examination in the UK would, by virtue of standardisation and increased sample sizes, enable a robust estimation of the correlation between medical school examination performance and subsequent performance in practice for UK doctors. But this would only show what we already know: higher performing medical students produce higher performing doctors on serially testing.

To build the specific evidence base for licensing examinations, regulators could make use of evaluative frameworks that explore *processes* or *outcomes*. Ultimately, approaches that involve a pre-post study design, ideally with a control group, are required. Without them we will not be able to truly understand if licensing examinations specifically provide a unique contribution to patient outcomes and safety above and beyond other forms of assessment and medical education.

1. Introduction

This report sets out our findings from a systematic review of the literature on licensing examinations for doctors and other healthcare professionals in the 49 countries currently comparable to the United Kingdom (UNDP, 2014)[1].

The review was carried out on behalf of the General Medical Council (GMC) and had three primary aims:

1. To establish the existing evidence base for the validity of medical licensing examinations or similar in countries comparable to the UK

2. To establish the validity evidence for the impact of medical licensing examinations

3. To identify best practice and any gaps in knowledge for medical licensing examinations.

The report provides details of the search strategies used and includes summaries of all the relevant literature located. It offers a critical synthesis of the literature, mapped to the American Psychological Association's (APA) validity framework (Downing, 2003), and an evaluation of the literature's quality and evidential value.

2. Background

The GMC maintains the health and safety of the public in the United Kingdom by regulating doctors at all stages of their training and subsequent professional development. Ultimately the GMC seeks to ensure all doctors licensed to practise in the UK meet the standards they set. Currently in relation to home graduates this is achieved by GMC setting standards for undergraduate education and undertaking inspections of medical schools. However UK graduates cannot ultimately be directly compared, for example, through the use of a national examination.

[1] The 49 comparable countries are those listed by the United Nations Development Programme (UNDP) in their most recent report on human development across the world. The UNDP calculates 'Human Development' by evaluating and assessing an index of component parts. These are: life expectancy at birth, mean years of schooling, expected years of schooling, gross national Income (GNI) per capita. The purpose of the Human Development Index is to measure *"average achievement in three basic dimensions of human development"*: a long and healthy life, knowledge, and a decent standard of living, see UNDP (2014) 'Human Development Report 2014 Sustaining Human Progress: Reducing Vulnerabilities and Building Resilience '. From this, countries are then ranked as having 'very high', 'high', 'medium', or 'low' human development - those countries in the 'very high' category include the UK.

In recent years, perhaps a more difficult challenge has been to regulate doctors trained outside the United Kingdom who come here to work. Although the migration of doctors is not a new phenomenon it is, for a variety of reasons, on the increase (Leitch & Dovey, 2010).

In Europe, for example, where the movement of doctors has a long history, agreements and directives exist to facilitate migration. The rights of citizens who live and work in the European Economic Area (EEA) to move and work freely across member states are enshrined in European law. As a member of the European Union (EU) and the EEA, the UK must embrace the concept of free movement between member states.

Although the principle of free movement within the EU and EEA provides medical regulators in Europe with some unique challenges, medical regulators across the world also appear to struggle with how best to regulate international medical graduates and other medical professionals who wish to move across national or regional (state) boundaries (Audas, 2005; Avery, Germano & Camune, 2010; Cooper, 2005; Doyle, 2010; Ferris, 2006; McGrath, Wong & Holewa, 2011).

In essence, the debate now centres on those who argue that patient safety, patient trust, and physician quality is best served through the use of licensure examinations (Melnick, 2009) and those who insist other methods of assessing physician competence are preferable (Ranney, 2006). Over time, the arguments employed to attack or defend these differing viewpoints have polarised. However, evidence to support these competing claims can be difficult to find. It is in this context that the GMC now wishes to know what evidence actually exists with respect to licensing examinations, and how this fits with the arguments for and against them.

For the purposes of this review a 'licensing examination' includes, but is not limited to, examinations that:

- Are taken close to the time of graduation
- Are set and administered at a national or regional level
- Cover generic skills
- Require success in the examination for a doctor to practice in the jurisdiction where the examination was taken.

In understanding what constitutes a national licensing examination a well-known and often cited example is the United States Medical Licensing Examination (USMLE). To practise medicine in the US, ***all*** medical doctors, whether trained in the US or elsewhere, must pass this three-step examination.

Whilst the USMLE is one example of a national licensing examination, many more countries operate licensing examinations that differ from this model. In some countries, medical graduates must pass a licensing examination as part of their medical training, and to work within the national jurisdiction. However, some graduates trained elsewhere who come to these jurisdictions find that their qualifications are considered equivalent to those of the home nation. Where these circumstances apply, there is no requirement for them to take the licensing examination to practise.

Equally, there are graduates trained in other parts of the world who find that their qualifications are not considered to have parity with those of the students in the home nation. As a result, these jurisdictions have devised one-off examinations to test the knowledge and competence of these graduates. Such examinations, though they do not require all doctors who work in the jurisdiction to take them, and are not necessarily taken at or near graduation, nonetheless serve a 'national' gate-keeping function and fit three of the four GMC criteria for a licensing or similar examination.

2.1 Methodology

To explore the topic and address the research aims we undertook a systematic review of the available literature – including grey literature[2]. This was supplemented by a website search of all those bodies with some involvement in regulating, licensing or otherwise authorising a medical practitioner's license to practise in the 49 countries comparable to the UK.

In many jurisdictions a combination of organisations play their part in the registration, regulation, and licensing of doctors. Some of the existing literature sets out who does what within these different systems, but many more remain opaque (de Vries, 2009; Kovacs et al., 2014; Rowe, 2005). Where the literature provided us with the sufficient information we sought out the relevant websites (where available).

We also surveyed as many of these regulatory or licensing bodies as we and the GMC had contact details for. The purpose of the survey was to gain additional literature on what informs their thinking on licensure examinations.

The predominant methodology throughout was a systematic review of the existing literature. Although systematic review is a continually evolving methodology, there exists a great deal of information on general protocols and what is currently regarded as 'best practice.' At all stages of the review (Grant & Booth, 2009; Popay, 2006) we adhered to this advice.

Best practice included setting out explicit criteria (see Table 1) for what literature would be included in the review and what was excluded (Bettany-Saltikov, 2010). It also required we establish a clear data extraction strategy (2009; Popay, 2006).

[2] Grey literature can include non-conventional print or electronic material not controlled by commercial publishers and can cover reports of different types, translations, theses, technical and commercial documentation etc. In this report it will also include material located on Internet websites.

Table 1: Inclusion and Exclusion Criteria

Inclusion Criteria	Exclusion Criteria
Medicine & Healthcare Professionals	Outside medicine
National or regional (State level)	Local or institutional level
Early Career/Graduation	Specialist examinations
Success in examination linked to ability to practise	Prior to 2005
Any language (assuming translation could be obtained)	
Countries comparable to the UK	
Published since 2005	

We devised our inclusion and exclusion criteria with the GMC's research questions in mind and the sort of evidence the GMC hoped to find. The information sought was extensive:

- The stated purpose or purposes of the exam
- Whether candidates are ranked to support recruitment into further training or employment
- The timing of national licensing exams
- The format
- The content
- Who takes the exam
- Details of ownership, funding, accountability and quality assurance
- Pass and failure rates
- Standard setting approaches
- Who the results are made available to
- Are the examinations used as a quality assurance mechanism for undergraduate medical education?
- Are there opportunities to retake the exam? If so what criteria attach this?
- Target and actual reliability values, validity and standard setting methods

We also wanted evidence for the impact of licensing examinations and whether they:

- Reduce variation in undergraduate curricula
- Drive ranking of organisations and individuals
- Increase confidence amongst employers, professional, and the public
- Lead to better skilled registrants
- Lead to higher standards of practice
- Emphasise knowledge and skills as opposed to values, behaviours and professionalism
- Promote cost effectiveness and high quality in summative examinations
- Have differential pass rates for graduates with different characteristics
- Predict candidates' subsequent performance in postgraduate training and practice
- Predict likelihood of referral to disciplinary proceedings

To assist us in this task, and in line with best practice, a minimum of two researchers were involved in the data extraction process with additional input from CAMERA's expert panel. The expert panel was a group of multi-disciplinary researchers with expertise in various areas of medical education, medical assessment methodologies, psychometrics, and statistics amongst other areas of expertise. This combination of knowledge and skill was valuable in dealing with the varied literature we identified and collected. It was also useful where differences of opinion were encountered.

Validity and the validity framework

A unique feature of this systematic review was the use of a 'validity framework' (Downing, 2003) developed by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME).

Assessment specialists recognise that validity is not a property of the test or assessment but of the meaning of the test scores (Messick, 1995). As Kane puts it *"it is only when an interpretation is assigned to the scores that the question of validity arises"* (Kane, Crooks & Cohen, 1999, p.6). To assist in the work of determining whether evidence and theory support the validity of test score interpretations, the *framework* identifies five distinct sources of validity evidence:

- Content
- Response process
- Internal structure
- Relationship to other variables
- Consequences

When looking for validity evidence the task is made easier by breaking the test down into these five areas, and considering how valid the evidence contained within each category is in supporting or refuting the test score interpretations (see Table 2).

In assessing validity, Downing also points out that:

> Some types of assessment demand a stronger emphasis on one or more sources of evidence … For example, a written, objectively scored test covering several weeks of instruction in microbiology, might emphasize content-related evidence, together with some evidence of response quality, internal structure and consequences, but very likely would not seek much or any evidence concerning relationship to other variables. (Downing, 2003, p.832)

In short, the validity framework classifies the sources of evidence that can potentially support test score interpretations into five broad areas:

Table 2: Summary of the Validity Framework

Content Evidence	The outlines, subject matter domains, and plan for the test as described in the test 'blueprint.' Mapping the test content to curriculum specifications and defined learning outcomes. The quality of the test questions and the methods of development and review used to ensure quality. The guidelines for scoring and administration. Expert input and judgements and how these are used to judge the representativeness of the content against the performance it is intended to measure.
Response Process	The clarity of the pre-test information given to candidates. The processes of test administration, scoring, and quality control. Evidence of candidate approaches to the test and what they try to do. The performance of judges and observers. Quality control and accuracy of final marks, scores, and grades.
Internal Structure	The statistical or psychometric characteristics of the test such as: • Item performance e.g. difficulty • Factor structure and internal consistency of subscales • Relationship between different parts of the test • Overall reliability and generalizability Matters relating to bias and fairness.
Relationship to other variables	The correlation or relationship of test scores to external variables such as: • Scores in similar assessments with which we might expect to find strong positive correlation. • Scores in related but dissimilar assessments e.g. a knowledge test and an Objective Structured Clinical Examination (OSCE) • Where weaker correlations might be expected. • Candidate factors such as age or level of training that might be associated with variation in test performance. Generalizability of evidence and limitations such as study design, range restriction, and sample bias.
Consequences	The intended or unintended consequences of the assessment on participants (such as failure) or wider societal impacts. The methods used to establish pass/fail scores. False positives and false negatives.

We used this framework to systematically organise the evidence found in the literature review and to structure our analysis and reporting.

Synthesis

The search elements of the review comprised of four phases: an initial scoping phase, the main review, an Internet search of the medical regulator/licensing authority sites of the 49 countries currently considered comparable with the UK, and a survey sent to medical regulators. As part of the main review, we also searched the websites of other organisations with a specific interest or professional stake in devising licensing examinations for healthcare professionals. These included the United States Medical Licencing Examination (USMLE) site and the numerous other sites affiliated to it such as the Educational Commission for Foreign Medical Graduates (ECFMG), the National Board of Medical

Examiners (NBME), Federal State Medical Board (FSMB), and the Foundation for Advancement of International Medical Education and Research (FAIMER).

Phase One

In the initial scoping phase it was important to test a broad range of databases to establish where most of the literature might be found. Prior knowledge and experience meant we were able to ultimately select seven databases as optimal:

- Embase (Ovid Medline)
- Medline (EBSCO)
- PubMed
- Wiley Online
- ScienceDirect
- PsychINFO
- BMJ

The chosen databases vary in size and subject content. In those with medical or healthcare profession sections such as EBSCO and EMBASE only these areas were searched.

We interrogated these databases using the search terms 'national licensing examinations for doctors' and 'national licensing exams for doctors.' Advanced search filters restricted the search to documents published between 2005 and 2014. Within these, and where advanced searching allowed it, the additional search terms 'dentists', 'nurses', 'midwives' and 'healthcare professionals' were applied. Search filters were set to find relevant material written in any language. Almost all the found literature was written and/or published in English.

After screening the search outputs via title and abstract, the scoping phase identified 128 potentially relevant documents. These papers offered a variety of qualitative, quantitative, and mixed methodologies together with editorials, opinion pieces, and personal views – mostly from acknowledged experts – across a range of healthcare professions.

Phase Two

Drawing on the experience gained in Phase One, the same seven databases were revisited for Phase Two. We used the same advanced search filters as the scoping phase so that no material prior to 2005 was retrieved. There was no restriction on language.

In debriefing and reflecting upon our experience of the scoping phase, CAMERA's expert panel suggested some changes to the inclusion criteria. As a result of these discussions we expanded our inclusion criteria and search strategy to include: 'International Medical Graduates', 'IMGs', 'International Medical Graduate programmes', 'International Medical Graduate examinations.'

During the scoping phase we also found that the terms ‘accreditation’, ‘credentialing’, ‘registration’, and ‘certification’ appeared in some of the material retrieved. Whilst ‘accreditation’ and ‘credentialing’ etc. are not synonyms for licensing or licensure, there are undoubtedly overlaps with licensure processes – this is especially so where ‘registration’ is concerned. To ensure our searches were thorough we included these four terms into our search strategy.

Our widened search strategy produced a large volume of search outputs but very modest returns. For example, an advanced search of EMBASE for International Medical Graduate ‘programmes’ and ‘examinations’, with filters, produced 1,895 hits. Screening the titles and abstracts of these reduced the number to 16.

In total, 87 potentially relevant papers were obtained.

Phase Three

The third phase of the review involved searching the Internet for the websites of medical regulators or those bodies with responsibility for licensing doctors and healthcare professionals in each of the 49 countries regarded as having ‘very high human development’ (UNDP, 2014). The purpose of these searches was to locate ‘grey’ literature relevant to the research objectives.

The amount of information on these websites varied considerably. The majority were in English or had an English version. Others were only accessible to English speakers via the ‘translate page’ function of various web browsers and search engines. The accessibility of these sites varied markedly. Sometimes, the only page accessible to English speakers was the ‘Home’ page.

In addition to the issue of accessibility, the task of searching these sites for relevant information and literature was complex. This was because of the way in which doctors and healthcare professionals are regulated, licensed, and registered and this varies from country to country (de Vries, 2009). Not all regulators, for example, have responsibility for licensing or registering doctors (Rowe, 2005). In some countries the licence to practice is the prerogative of the Ministry of Health, while elsewhere, it belongs to a regional or professional body or a combination of the two (Rowe, 2005).

In other jurisdictions, the granting of a licence to practise may be only one step in a complex process. In these instances it was necessary to locate the websites of the other parties involved to see what if any literature and information they could provide. Once again, the level of accessibility and the quality of information available varied widely (de Vries, 2009).

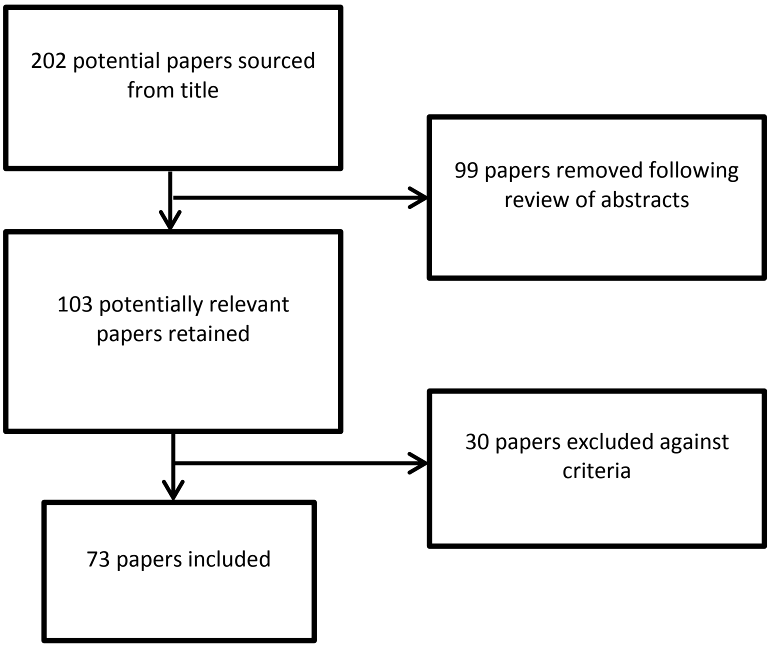
This phase of the review also included a search of websites belonging to assessment specialists specifically, but not exclusively, USMLE and its various partner organisations ECFMG, NBME, the FSMB, and FAIMER. These sites were very accessible and contained a great deal of information on the range of services and products.

Our searches in this phase of the review were aided by grey literature research that compared or examined medical regulation and healthcare in various countries across Europe and the world (de Vries, 2009). Furthermore, the use of general Internet searches was also useful. Together, the information from these sources directed us to a number of websites that offered useful information to healthcare professionals interested in working in different countries. Broad Internet searches also led us to blogs and other social media offering anecdotal advice, sometimes supplemented with website details. Overall, this aspect of the review yielded 14 potentially relevant documents.

During the three phases of the review 202 documents were downloaded. After more detailed screening of the titles and abstracts against the inclusion and exclusion criteria this number was reduced to 103.

One researcher reviewed the full text of all 103 papers, more than half of these were reviewed by other members of the research team. Through this process 30 papers were excluded for not meeting the inclusion and exclusion criteria. The total number of papers included in the final review was 73 (see figure 1).

Figure 1: Overview of the literature search process


202 potential papers sourced from title
99 papers removed following review of abstracts
103 potentially relevant papers retained
30 papers excluded against criteria
73 papers included

Data extraction from the included papers was shared between team members in a similar way to the reading/screening process.

Phase Four

Finally, as an adjunct to the search process we collaborated with the GMC to devise an online survey. This was sent, by the GMC and the International Association of Medical Regulatory Authorities (IAMRA), to medical regulators and licensing authorities they had contact details for in the 49 countries comparable to the UK. Our intention with the survey was to elicit information on any other grey or unpublished literature used by regulators to inform their thinking on licensure examinations. Excluding duplicates we received 11 replies. These contained 3 references to literature, which we had already obtained through our searches but no other additional manuals, documents or information sources.

3. Findings

The final papers in the review had a mix of methodologies, explore different aspects of the licensing process, and advocate a variety of viewpoints.

After being mapped to the APA validity framework, only 23 of the 73 papers were found to contain validity evidence for licensing examinations. The remaining 50 papers, some of which are important in terms of shaping the arguments, consisted of informed opinion and editorials, or simply described and contributed to the continuing debate.

In this review we first summarise the 50 papers to offer some context to the national licensing examination debate. Once the landscape of licensure is set out, we then describe and evaluate the core papers in more detail in order to address the three primary aims of the research and provide answers to at least some of the GMC's questions.

3.1 The landscape of licensure

The literature on medical licensing regimes across the world is reasonably extensive. But, it is also far from complete (de Vries, 2009; Kovacs *et al.*, 2014; Rowe, 2005). The literature that does exist is of sufficient scope and quality to capture some of the differences and similarities between the diverse regulatory and licensing regimes that exist within some of the 49 jurisdictions.

The literature indicates that four different approaches to licensing examinations exist. These may be summarised as:

1. Where student doctors trained in the national jurisdiction (home students) must pass a national licensing exam as part of their medical study and to obtain a licence to practise
2. Where ***all*** prospective doctors must pass a national licensing examination to obtain a licence to practise within the national jurisdiction
3. Where international medical graduates must pass an examination where their qualifications are not recognised as compatible with those of students trained in the national jurisdiction
4. Where no national licensing examinations exist

The countries where the literature shows this first approach exists are Germany (Seyfarth et al., 2010), Switzerland, Poland (http://www.cem.edu.pl/), Bahrain (www.moh.gov.bh/PDF), Qatar, and Croatia (www.qchp.org.qa, www.hlk.hr/MedicalLicence). In these national jurisdictions all home trained students must pass the examination to obtain a licence to practise, but exemptions exist for some international medical graduates. For example in those countries that are part of the EU/EEA, graduates from other EU/EEA countries are exempt.

The second approach to national licensing examinations requires that all prospective doctors who aspire to practise medicine within the jurisdiction must, regardless of where they are trained, pass the licensing examination. There are no exemptions. The literature around this form of examination, has predominately developed in North America (Sutherland, 2006). Of the 49 comparable countries only the US, Canada, Hong Kong, Japan, Korea, Chile, and the United Arab Emirates (UAE) use this approach to licensing.

The amount of information available about these examinations, their content and their quality, varies. The North American literature is considerable and reasonably broad in its subject matter. The information provided by the UAE is also reasonably extensive. For Korea, the information is more limited and comes mainly from two papers on the Korean examination. These papers provide detail on the structure of the exam, and academic arguments for changing the cut scores (Ahn & Ahn, 2007; Lee, 2008). In contrast, the Medical Council of Hong Kong provides only brief descriptive detail on the number and type of questions in their three part examination www.mchk.org.hk, while Japan provides virtually nothing. Table 3 summarises the available information.

Although the system of licensing in these countries would appear to eliminate any ambiguity over eligibility, the literature also highlights that the *post* examination road to practice is not necessarily more straightforward. From the perspective of those who pass but less well – i.e. those in the lowest quartile - whether they are home graduates or from elsewhere, the use of applicant ranking can have implications for future career opportunities: in short, there is evidence to suggest that those who get the highest scores are likely to get the best jobs (Green, Jones & Thomas, 2009; Kenny, McInnes & Singh, 2013; Noble, 2008). It can also have an impact on the health of graduates, many of whom complain of ‘stress and burnout’ in striving to achieve the best grades (McMahon & Tallia, 2010).

Table 3: National Licensing Examinations and Component Parts

Country & Examination	Component parts				
	Part 1	Part 2	Part 3	Pass mark	Candidates
Australia	The AMC Computer Adaptive Test (CAT). 150 'A-Type' MCQs (one correct response from five options). 120 scored items, 30 non-scored pilot items. Candidates are expected to complete all 150 MCQs. Tests knowledge of the principles and practice of medicine in general practice, internal medicine, paediatrics, surgery, obstetrics & gynaecology. Candidates must pass this examination to go to take the AMC Clinical Examination.	AMC Clinical Examination: assesses clinical skills in medicine, surgery, obstetrics, gynaecology, paediatrics, and psychiatry. Also assesses ability to communicate with patients, their families and other health workers. OSCE style, 16 component multi-station assessment.	N/A	Pass scores in 12 or more stations including at least a pass in obstetrics/gynaecology and one pass in paediatrics qualifies candidate for AMC certificate. Pass scores in 10 or 11 stations including at least one pass in obstetrics/gynaecology and one pass in paediatrics leads to a retest. Pass scores in 9 stations or less or fails in all three obstetrics/gynaecology stations or fails in all three paediatrics stations is regarded as a clear fail	These are examinations on the Standard Pathway i.e. the pathway for those IMGs who do not qualify for the other pathways into the Australian workforce.
Bahrain BMLE	Written test 100 MCQs starts with a stem followed by 4 or 5 responses. Only one is correct	2nd part of Part 1. A written test uses MCQs to assess 10-15 Patient Management Problems. Assesses clinical reasoning skill & ability	OSCE of 20 clinical stations. Each scenario followed by 3-5 questions.	50% or above for written to be eligible for OSCE. Final passing score cumulative of 60% or above from written & OSCE	Taken by all doctors who wish to practise in Bahrain. No detail on retake restrictions.

Country & Examination	Component parts				
	Part 1	Part 2	Part 3	Pass mark	Candidates
Canada MCCQE	One day computer based test in two parts. Morning session: 196 MCQs, afternoon session Clinical Decision Making component – short menu, short answer questions.	Assesses competence, specifically knowledge, skills, & attitudes using OCSE style simulation stations.	N/A	Determined by the Central Examination Committee	IMGs and International medical students (IMS)s must pass the Medical Council of Canada Evaluating Examination (MCCEE).
Chile: EUNACOM	Written test with 180 MCQs (two sections of 90 questions each in 7 thematic areas)	Practical examination of general practice. Clinical evaluation in a real or simulated environment in the areas of medicine, surgery, obstetrics- gynaecology & paediatrics	N/A	Minimum score defined by Ministry of Health.	Taken by all doctors who wish to practice in Chile. http://www.eunacom.cl/
Croatia: Croatian Medical Licensing Examination	No detail available	No detail available	No detail available	No detail available	Taken by Croatian Graduates and non EU/EEA nationals. EU/EEA nationals are exempt.
Finland: Professional Competence Examination	Written examination on key areas of medicine	Written examination on healthcare management	Oral examination in a clinical setting (with patient present)	No detail available	For non EU/EEA nationals. Graduates do not need to provide proof of language ability to be licensed.
France: Epreuves Classantes Nationales NCE (ranking examination).	Theory test	No detail available	No detail available	Pass mark required	N/A
Germany: *Staatsexamen*	M1 *Physikum* or preclinical medicine after 2 years.	N/A	M2 written and oral practical includes the 'Jawbreaker' – includes the content of the entire clinical phase. MCQs.	No detail available	Only German doctors take these examinations. Non EU/EEA nationals may be required to take a 'knowledge test' to prove their qualifications are equivalent to German standards (Chenot, 2009)
Hong Kong: The Licensing Examination.	Examination in Professional Knowledge 120 MCQs to test knowledge in basic science, medical ethics, community medicine, medicine, surgery, orthopaedic surgery, psychiatry, paediatrics & obstetrics & gynaecology	Proficiency Test in Medical English (scheduled for March 2015).	The Clinical Examination: to test how candidates apply professional knowledge to clinical problems. (scheduled for May/June 2015)	No detail available	All IMGs must pass Parts 1 & 2 to take Part 3. No retakes until appeal process verdict

Country & Examination	Component parts				
	Part 1	Part 2	Part 3	Pass mark	Candidates
Ireland: Pre Registration Examination System (PRES)	All applicants undergo Level 1 assessment and verification of their documentation. Those not exempt after this process go on to take the next parts.	Level 2: Computer-based examination using MCQs. Pass is required to move to level 3. Level 2 pass is valid for 2 years.	Level 3: Assessment of Clinical Skills. OSCE style examination. Interpretation Skills test is one paper based examination. Level 3 must be taken within 2 years of passing level 2. Level 3 pass	No detail on level 2 pass marks. Level 3: each station/question is marked out of a total of 20. Each skills component is marked out of 120 marks.	Non EU/EEA graduates may be required to take a Medical Council Examination unless exempt. Three attempts at level 2 & 3 are allowed.
Israel	Written examination in Hebrew uses MCQs	N/A	N/A	No detail available	Home and IMGs. Passing score is valid for 3 years. http://www.ima.org.il/ENG/Default.aspx
Japan: National Medical Licensing Examination (NMLE)	No detail available	N/A	N/A	Not available	Taken by all those who wish to work in Japan. Test is in Japanese. http://www.med.or.jp/english/
Korea: KMLE	Written examination	Clinical Skills test OSCE style.	N/A	Candidates who score 60% in the written test with at least 40% in each subject are deemed successful. Part 2 pass scores are determined by the deliberation committee of medical school professors	Overseas qualifications must be recognised by the Minister of Health & Welfare prior to IMGs taking the test.
New Zealand: NZREX (Clinical)	OSCE consists of 16 stations covered. Competencies tested: History taking, clinical examination, investigating, management, clinical reasoning. Also, communication and professionalism will be assessed.	N/A	N/A	Criterion referenced, contrasting groups system to determine pass score.	No limit to how many times it can be taken. Eligibility requirements must be satisfied on each occasion. IMGs only

Country & Examination	Component parts				
	Part 1	Part 2	Part 3	Pass mark	Candidates
Poland: State Physician & Dental Final Exam	The SP/DE is a written test, in Polish and consists of 200 MCQs only one correct answer out of the choices. Mix of medical knowledge, questions about specific medical processes, analysis of medical records, and establishing medical diagnosis.	No practical examination at present.	N/A	Pass/Fail threshold 54%	Content of the examination does not exceed the scope of the internship programme. Oral skills are not tested. Medical schools test communication and procedural competencies.(Nowakowski, 2013) Taken by IMG and not EEA candidates
Portugal: 'Exame Nacional de Seriacao' (Ranking examination for residency posts)	Written test MCQs on internal medicine.	N/A	N/A	Detail not available	(Pavão Martins, 2013)
Spain: MIR (National Residency Examination) 'examen MIR'	Written test, 250 MCQs	Clinical competence, no other detail.	N/A	Not available	Used for ranking - places allocated on the basis of MIR exam and an evaluation of candidate's academic record. No limit to the times it can be taken.(Lopez-Valcarcel et al., 2013)
Sweden: TULE-test	Written test of medical knowledge. 100 questions	Practical tests over 2 days	N/A	65% for medicine & surgery	TULE is for IMGs qualified outside the EU/EEA/Switzerland. 3 retakes allowed. Can re-test parts of Part 1 failed. Must retake both parts of Part 2
Switzerland: Federal Licensing Examination: FLE	Locally administered written examination using MCQs	Clinical Skills OSCE style examination.		No detail available.	Swiss graduates must take the FLE. Non EU/EEA graduate qualifications are assessed at Cantonal level. IMGs take the test to if they wish to practise independently.
UaE	No detail available	No detail available	No detail available	No detail available	3 retakes allowed. Separate registration required to work in Dubai.

Country & Examination	Component parts				
	Part 1	Part 2	Part 3	Pass mark	Candidates
United States USMLE	Step 1: 322 MCQs to test and measure basic science knowledge. Consists of 7 blocks of 46 items. 1 hour for each block of test items. Maximum of 7 hours testing.	Step 2 consists of 2 components: Clinical Knowledge test. 350 MCQs divided into 8 blocks. 1 hour for each block. The number of questions per block varies but does not exceed 44. Clinical Skills test: 12 OSCE style patient encounters using standardised patients.	Step 3: 2 day examination. First day has 256 MCQs divided into 6 blocks of 42-43 items. 1 hour per block. Second day: 198 MCQs divided into 6 blocks of 33 items. Includes a 7 minute Computer-based Case Simulation (CCS) tutorial, then 13 case simulations.	Step 2 CS is a pass/fail examination. Current minimum pass scores: Step 1: 192 Step 2 CK: 209 Step 3: 190	Very detailed information on all parts of the USMLE on their site. IMGs must be certified by the Educational Commission for Foreign Graduates (ECFMG0) to take USMLE Step 3
Qatar: Qualifying Examination	No detail available	No detail available	No detail available	No detail available	International Medical Graduates not exempt.

The third form of licensing examination differs from the first two in that some may not perceive it to be a truly 'national' licensing examination; as they only target IMGs. Nevertheless, these examinations fit three of the four elements of the GMC definition for a licensing examination in that they are:

- Set and administered at a national level
- Cover generic skills
- Success in the exam is necessary to practise as a doctor in the jurisdiction where the exam is taken.

In countries like Australia and New Zealand[3] detailed information is available on medical council websites (www.amc.org.au; www.mcnz.org.nz) to allow prospective doctors to determine what 'pathway' into the physician workforce their qualifications require them to follow. Some international medical graduate qualifications are considered to have parity with those of Australasian medical graduates, but many others are not. The process of establishing which qualifications are acceptable and which are not is straightforward in these two jurisdictions. In Europe, where EU and EEA member countries are bound by directives that allow the free movement of citizens across member states, this type of licensing examination is limited to those who come from outside the EU/EEA (Kovacs *et al.*, 2014).

For non-EU/EEA graduates this means that across Europe this group of doctors will participate in a plethora of examination processes to gain entry to the profession in their chosen jurisdiction. Some international medical graduates (and those who research them) suggest these processes are flawed. There is certainly little specific information in many member states about the examination itself or the process of which it is part.

In Sweden non-EU/EEA graduates describe the process in that country as disorganised, bureaucratic, and more strict than the process for EU/EEA graduates (Musoke, 2012). Other research studies indicate the same happens elsewhere in Europe (Sonderen *et al.*, 2009). Where this sort of examination approach is concerned, and unlike the previous approach, it is important to stress that there is no large and easily accessible body of research data from which we can draw.

Although all three approaches vary in the degree of openness and clarity that surrounds the examination process, some element of pragmatism is discernible. The demands that come from a widespread shortage of physicians across jurisdictions (Leitch & Dovey, 2010) mean that international medical graduates who have not passed the requisite examinations in

[3] Australia & New Zealand, like other countries including the UK, operate an 'accreditation' model of licensing regulation. In each case IMGs are required to provide evidence of language competence and validated documentation of their primary qualifications.

their chosen jurisdictions, are not always entirely disbarred from working in them. Certainly within North America, where the second form of national licensing operates, an extensive support system exists to help international medical graduates prepare for the licensing examinations they must take (Audas, 2005; Maudsley, 2008).

Finally, the literature also reveals that other jurisdictions such as Malta and Kuwait amongst others, eschew the use of national licensing examinations to assess the eligibility of a medical practitioner's qualifications and the granting of a license. The absence of national licensing examinations is no less significant than their presence.

These different approaches to regulation and licensing are, of course, the result of the historic, cultural, economic, political, and geographic contexts in which these systems have evolved and currently exist (Borow, Levi & Glekin, 2013). Therefore, before we examine the current debate on the different approaches to national licensing, we should briefly survey some of the other features of what is a diverse landscape of medical regulation to give some context.

Academic debate around licensure examinations

The debate around one-off national licensing examinations is emotive and often polarised (Ferris, 2006; Neilson, 2008). The arguments for and against revolve around a limited number of core themes (Harden, 2009; Melnick, 2009; Ricketts & Archer, 2008). Because the 'evidence' for these is, in the main, equivocal and rhetorical, supporters and opponents often cite the same research to warrant or back their argumentative stance.

Supporters of North American style licensing examinations argue an examination that all graduates must take and pass to enter their chosen profession is 'fair' (Melnick, 2009). Certainly, by virtue of the fact that everyone must take it, claims that such tests discriminate or favour is lessened (Lee, 2008). However, as the statistical evidence demonstrates, graduates brought up within the home nation's medical education system and who are familiar with the language, appear to have some advantage over those trained elsewhere (Holtzman et al., 2014; Tiffin et al., 2014).

In the same argumentative vein, supporters suggest the logistics of running these examinations necessarily requires resources and expertise to be pooled. The financial costs are, as even supporters concede, high (Lehman & Guercio, 2013). The result however is that the quality of these examinations is also said to be high (Lehman & Guercio, 2013; Neumann & Macneil, 2007). This, it is argued, ensures a minimum standard or benchmark is set for all those entering the profession and the process is standardised (Cosby Jr, 2006). Proponents also claim that because the quality of prospective entrants to the profession has been assessed in this way for a long time, the longevity of the process is itself an endorsement of its integrity (Melnick, 2009; Neumann & Macneil, 2007).

There is little doubt that the technical quality of the North American style assessments is supported by good empirical evidence (CUP, 2008; Guttormsen et al., 2013; Hecker & Violato, 2008; Lillis, 2012; Margolis *et al.*, 2010; Stewart, Bates & Smith, 2005). And while educational assessment and theory continues to evolve at a rapid pace, it is not unreasonable to regard the USMLE and Medical Council of Canada Qualifying Examination (MCCQE) as being the 'gold standard' of licensing examinations (Bajammal et al., 2008; Lillis, 2012; Norcini *et al.*, 2014). The difficulties arise however from the additional claims premised on the back of the quality assessment evidence – most notably the patient-safety/public trust arguments – that good assessment directly leads to better patient care (McMahon & Tallia, 2010; Stewart, Bates & Smith, 2005; Tamblyn *et al.*, 2007; Wenghofer *et al.*, 2009).

The arguments that North American style licensing examinations make patients safer rest on assumptions rather than on evidence (Harden, 2009). The first assumption is that the public are safer because assessment specialists who devise and oversee the USMLE provide a credible 'external audit' of the quality of medical graduates. Second, that assessment specialists are able to accurately recognise what constitutes a minimum standard of knowledge and competence and accurately assess it (Melnick, 2009). Third, that a statistical correlation between examination scores, patient care outcomes, and disciplinary action in later professional life is evidence of a causal relationship (Tamblyn *et al.*, 2007; Wenghofer *et al.*, 2009). In contrast to the evidence for quality of the assessment themselves, the empirical evidence to back these broader claims is sparse (Boulet & van Zanten, 2014; Sutherland, 2006).

Opponents of national licensing examinations argue against national licensing examinations on a number of grounds (Harden, 2009). Some, particularly US dentists, point out that when licensing examinations were introduced the social, educational, and professional context was different:

> We see a system designed over 100 years ago to solve a problem that no longer exists – proprietary diploma mills that had no educational standards, or accreditation. (Ferris, 2006, p.129)

Advances in medical education and in understanding how students learn, now point to other ways of assessing competence and skill. Opponents of licensing examinations advocate alternatives such as the New York Dental Residency programme (Ferris, 2006). They do so on the basis that "*the best preparation for the practice of dentistry is the practice of dentistry*" (Calnon, 2006, p.140).

Standardisation in medical education is also viewed by opponents as a cause for concern – primarily, but not entirely - amongst critics outside of North America (van der Vleuten, 2009). In the same way as supporters for licensure examinations employ a 'why-change-what-isn't-broken' narrative, so opponents in the UK and Europe employ the same narrative to defend the virtues of a diverse medical curricula (Gorsira, 2009; Harden, 2009).

A further criticism is that a 'one-point-in-time' early career licensing examination is not an effective way to measure physician competence or to anticipate later behaviours and professional practice. Opponents of licensing examinations argue that the more easily accessible 'learning outcomes' are what get tested and not those related to overall competence (Harden, 2009; Neilson, 2008; Noble, 2008). A continuing programme of on-the-job assessment, appraisal, and professional development they suggest, provides more accurate and up-to-date evidence of practitioner competence (Calnon, 2006; Kovacs *et al.*, 2014; Waldman & Truhlar, 2013). A system already well established in the UK.

Neilson following the same argumentative line as Calnon (2006) gives an experienced practitioner's view:

> The standardisation of final licensing and fitness to practise examinations may make educationalists weep with joy, but there is no clear evidence that it makes for better doctors. My colleagues and I deal with the immediate postgraduate training of juniors and know that, regardless of where the doctors have qualified, their practical education starts when they start working with patients for real. (Neilson, 2008)

Evidence that those who score highest in early career examinations go on to get the best jobs (Green, Jones & Thomas, 2009; Kenny, McInnes & Singh, 2013) - also supports the 'predictive' ability of these examinations (Harden, 2009).

Harden cites a number of studies and alternative approaches to education and learning, to argue that trying to predict which doctors may appear in disciplinary hearings or administer poor patient care is subject to a myriad of other variables and consequences. And, as those researchers who make the claims themselves point out (Norcini *et al.*, 2014; Tamblyn *et al.*, 2007; Wenghofer *et al.*, 2009), those doctors they identify as more likely to be disciplined still passed the examination.

From a European perspective, opponents of North American style licensing examinations argue the practicalities of introducing such an examination across Europe are considerable, given the unique mobility arrangements that exist in this region. As we noted earlier, medical regulators in the EU and EEA must abide by imperatives that ensure EU and EEA citizens are able to freely move and work across jurisdictions (de Vries, 2009). Securing consensus and then devising a mechanism whereby *all* medical doctors in Europe sit a national or European licensing examination is seen as difficult (Gorsira, 2009).

Van de Vleuten carefully weighs the advantages and disadvantages of a 'pan-European' licensing exam. He takes an ethical and pragmatic view:

> My personal view is that there is no escaping the argument that the public is entitled to this reassurance, particularly in the open professional community across Europe. That is why we need to start thinking very carefully about how qualifying systems could be set up to achieve the desired effects without doing too much harm to learning and to innovation power. A first step has been taken in the Netherlands, where we have set up a collaboration of medical schools in developing and administering progress tests across five of the eight medical schools … In this case we can speak of a

> fully bottom-up process towards a near national exam that is completely governed by the participating medical schools. I am aware that this model may not work in other European countries … Taking a European perspective in such a development seems much more desirable, albeit complicated, than reinventing the wheel at all the national levels. (van der Vleuten, 2009, p.191)

Finally, a large body of predominantly US literature relating to other healthcare professions suggests national licensing examinations have only a limited use within the complex system of regulation that exists in North America. The experiences of these professionals indicate such examinations do little to assist in increasing the mobility of health professionals (Cooper, 2005; Philipsen & Haynes, 2007). It seems additional layers of intra and inter-state regulation involving certification, credentialing, and accreditation interwoven with regulatory politics make for a confusing and obstacle-ridden landscape that does little to make things clear for practitioners or public alike (Rehm & DeMers, 2006).

The evidence for validity

As we have seen so far there continues to be lengthy debate about the value of national licensing examinations. While there is much debate, empirical evidence to support the arguments for or against is less forthcoming. In this last section we present and critique the 23 key papers that attempt to provide validity evidence *for* licensing examinations. To do this we have mapped the 23 papers to the APA validity framework and then summarised the analysis. The 23 papers are listed in Table 4.

Table 4: Papers providing empirical evidence for the validity of licensing examinations

Content	Response process	Internal structure	Relationship to other variables	Consequences
CEUP (2008): ‘Comprehensive Review of USMLE Summary of the Final Report and Recommendations’ Lillis, S., Stuart, M., Sidonie, Takai, N. (2012): ‘New Zealand Registration Examination (NZREX Clinical): 6 years of experience as an Objective Structured Clinical Examination (OSCE)’ Ranney, R.R. (2006): ‘What the Available Evidence on Clinical Licensure Exams Shows.’ Guttormsen, S., Beyeler, C., Bonvin, R., Feller, S., Schirlo, C., Schnabel, K., Schurter, T., Berendonk, C. (2013): ‘The new licensing examination for human medicine: from concept to implementation.’	Lillis, S., Stuart, M., Sidonie, Takai, N. (2012): ‘New Zealand Registration Examination (NZREX Clinical): 6 years of experience as an Objective Structured Clinical Examination (OSCE)’ Seyfarth et al., (2010): ‘Grades on the Second Medical Licensing Examination in Germany Before and After the Licensing Reform of 2002.’ Guttormsen, S., Beyeler, C., Bonvin, R., Feller, S., Schirlo, C., Schnabel, K., Schurter, T., Berendonk, C. (2013): ‘The new licensing examination for human medicine: from concept to implementation.’	Harik, P., Clauser, B.E., Grabovsky, I., Margolis, M.J., Dillion, G.F., Boulet, J.(2006): ‘Relationships among subcomponents of the USMLE Step 2 Clinical Skills examination, the Step 1, and the Step 2 Clinical Knowledge examinations. Lillis, S., Stuart, M., Sidonie, Takai, N. (2012): ‘New Zealand Registration Examination (NZREX Clinical): 6 years of experience as an Objective Structured Clinical Examination (OSCE)’ Ranney, R.R. (2006): ‘What the Available Evidence on Clinical Licensure Exams Shows.’ Guttormsen, S., Beyeler, C., Bonvin, R., Feller, S., Schirlo, C., Schnabel, K., Schurter, T., Berendonk, C. (2013): ‘The new licensing examination for human medicine: from concept to implementation.’	Cuddy, M.M., Dillion, G.F., Holtman, M.C., Clauser, B. (2006): ‘A Multilevel Analysis of the Relationships Between Selected Examinee Characteristics and United States Medical Licensing Examination Step 2 Clinical Knowledge Performance: Revisiting Old Findings and Asking New Questions.’ Harik, P., Clauser, B.E., Grabovsky, I., Margolis, M.J., Dillion, G.F., Boulet, J.(2006): ‘Relationships among subcomponents of the USMLE Step 2 Clinical Skills examination, the Step 1, and the Step 2 Clinical Knowledge examinations.’ Hecker K, & Violato, C. (2008): ‘How much do differences in Medical Schools Influence Student Performance? A Longitudinal Study Employing Hierarchical Linear Modelling. Kenny, S., McInnes, M., Singh, V. (2013): ‘Associations between residency selection strategies and doctor performance: a meta analysis.’ McManus, I., & Wakeford, R. (2014): ‘PLAB and UK graduates performance on MRCP(UK) and MRCGP examinations: data linkage study.’ Ranney, R.R. (2006): ‘What the Available Evidence on Clinical Licensure Exams Shows.’ Stewart, et al., (2005): ‘Relationship Between Performance in Dental School and Performance on a Dental Licensure Examination: An Eight Year Study.’ Tiffin et al., (2014): ‘Annual Review of Competence Progression ARCP Performance of doctors who passed Professional and Linguistic Assessments Board (PLAB) tests compared with UK graduates.’ Zahn et al., (2012): ‘Correlation of National Board of Medical Examiner’s Scores with the USMLE Step 1 and Step 2 Scores.’ Guttormsen, S., Beyeler, C., Bonvin, R., Feller, S., Schirlo, C., Schnabel, K., Schurter, T., Berendonk, C. (2013): ‘The new licensing examination for human medicine: from concept to implementation.’	Ahn, D., & Ahn, S. (2007): Reconsidering the Cut Score of the Korean National Medical Licensing Examination Green, M., Jones, P., Thomas Jr, J.X. (2009): ‘Selection Criteria for Residency: Results of a National Program Directors Survey.’ Holtzman et al., (2014): ‘International variation in performance by clinical discipline and task on the United States Medical Licensing Examination Step 2 Clinical Knowledge Component.’ Kenny, S., McInnes, M., Singh, V. (2013): ‘Associations between residency selection strategies and doctor performance: a meta analysis.’ Kugler, A. D, & Sauer, R.M. (2005): Doctors without Borders? Relicensing Requirements and Negative Selection in the Market for Physicians.’ Lillis, S., Stuart, M., Sidonie, Takai, N. (2012): ‘New Zealand Registration Examination (NZREX Clinical): 6 years of experience as an Objective Structured Clinical Examination (OSCE)’ Margolis et al., (2010): ‘Validity Evidence for USMLE Examination Cut Scores: Results of a Large Scale Survey’ Musoke, S. (2012): ‘Foreign Doctors and the Road to a Swedish Medical License.’ Norcini et al., (2014): ‘The relationship between licensing examination performance and the outcomes of care by international medical school graduates.’ Ranney, R.R. (2006): ‘What the Available Evidence on Clinical Licensure Exams Shows.’ Stewart, et al., (2005): ‘Relationship Between Performance in Dental School and Performance on a Dental Licensure Examination: An Eight Year Study.’ Sutherland, K., & Leatherman, S. (2006): ‘Regulation and Quality Improvement A Review of the Evidence.’ Tamblyn et el., (2007): ‘Physician Scores on a National Clinical Skills Examination as Predictors of Complaints to Medical Regulatory Authorities.’ Wenghofer et al., (2009): ‘Doctors Scores on National Qualifying Examinations Predict Quality of Care in Future Practice.’ Guttormsen, S., Beyeler, C., Bonvin, R., Feller, S., Schirlo, C., Schnabel, K., Schurter, T., Berendonk, C. (2013): ‘The new licensing examination for human medicine: from concept to implementation.’

Content validity

Content validity includes the outline and plan for the test. The principal question to ask is whether the content of the test is sufficiently similar to and representative of the activity or performance it is intended to measure?

Four papers offer some evidence for the content validity of specific licensing examinations:

Lillis, S., Stuart, M., Sidonie, Takai, N. (2012): 'New Zealand Registration Examination (NZREX Clinical): 6 years of experience as an Objective Structured Clinical Examination (OSCE)' argues that the NZREX OSCE, an examination that is comprised of a series of simulations of lived-world activities, is both valid and educationally robust. In designing these simulations and as part of constructing a standardised and auditable approach a blueprint is devised. The paper describes some of the blueprint material and how this has evolved and altered over a 6-year period to improve the quality of the simulations. The research on which the examination rests involved a literature review, expert opinion in the form of a working group, and assessment of previous incarnations of the examination against what is regarded as 'best practice.' The NZREX Clinical is an important examination because it provides a pathway to practice for IMGs who do not meet the requirements of other pathways into the New Zealand medical profession. The paper thus puts the examination into context. It describes the OSCE in detail e.g., *"NZREX Clinical is an OSCE format of 16 stations. Each station lasts for 12 minutes …"* The paper sets out what sort of knowledge and skills are being tested i.e., 'medical' or 'surgical', the statistical methods used as part of a continuing quality control process, that professional actors are used in the role play, the extent of their training, the location of examination, and so on. In so doing the authors argue for the validity of the assessment by providing evidence for its construction and rigorous blueprinting to clinical domains and clinical reasoning skills. The paper provides a good overview of the importance of the blueprint process in designing examinations.

CEUP (2008): 'Comprehensive Review of USMLE Summary of the Final Report and Recommendations' was a review of the USMLE in 2008 to *"determine if the mission and purpose of USMLE were effectively and efficiently supported by current design, and the format of the USMLE. This process to be guided, in part, by an analysis of information gathered from stakeholders, and was to result in recommendations to USMLE governance"* (p1). A committee consisting of 19 members approved by the CEOs of the NBME and FSMB, with two thirds that had *"… direct experience with the USMLE program and about one third did not"* (p5) concluded that USMLE was not 'broken.' They surveyed the stakeholders including the public and held 27 stakeholder meetings. The data revealed several *"general trends"* and they made 6 recommendations:

1. To design a series of assessments to support decisions about a physician's readiness to provide patient care at the interface between undergraduate and graduate practice, and at the beginning of independent practice.
2. To adopt a general competencies schema for the design, development, and scoring of USMLE and a research agenda to find new ways to measure general competencies.

3. To emphasise the scientific foundations of medicine in all components of the assessment process.
4. The assessment of clinical skills to remain a component of the USMLE, but to consider ways to enhance the test methods used.
5. To introduce a test format to assess examinees' ability to recognise and define a clinical problem and their ability to find scientific and clinical information to address the problem.
6. USMLE to encourage the NBME to meet the 'assessment needs' of secondary users of USMLE.

Ranney, R.R. (2006): 'What the Available Evidence on Clinical Licensure Exams Shows' examines the evidence for the reliability, content validity, and concurrent validity of initial licensure examinations in US dentistry. The paper uses a traditional narrative literature review to gather information and evidence. Where evidence for content is concerned, the author notes an absence of adequate evidence.

Guttormsen, S., Beyeler, C., Bonvin, R., Feller, S., Schirlo, C., Schnabel, K., Schurter, T., Berendonk, C. (2013): 'The new licensing examination for human medicine: from concept to implementation.' The authors set out the development of the Swiss Federal Licensing Examination. They discuss in thorough detail the development and piloting of the examination content for the written (MCQ) and clinical skills components. They do not provide an empirical evaluation of the process or the quality of the results. Also, there are no sample questions or stations. The examination blueprint is also not presented, although experts from the US and Canada were used to guide parts of the process. The data reported is from one year of the examination (785 candidates). The study is observational and descriptive with some quantitative analysis.

Analysis

The inclusion criteria for our review (no material prior to 2005 to be included) meant there is limited published material on content validity for national licensing examinations. The Lillis et al. (2012) paper represents a good example of the challenge of finding content validity evidence for licensing examinations. The authors present and describe evidence for the validity of the examination they have constructed and how it compares to other licensing examinations, but fail to critically appraise it.

The USMLE report CEUP (2008) is a précis of the comprehensive review. It describes the review, the rationale for the review, and the recommendations in generalised descriptive terms. Although the report demonstrates USMLE's commitment to product improvement, the review provides no technical detail. We had hoped to identify technical detail on the USMLE such as blueprinting exercises or an 'assessment manual' through the survey or our online searches. Whilst USMLE and the other organisations associated with it have a substantial online presence, specific detail on their products is (presumably for commercial reasons) not freely available - although excellent guidance about the process is available for prospective candidates, http://www.usmle.org/pdfs/bulletin/2015bulletin.pdf.

The paper from Guttormsen et al., (2013) describes in close detail the process by which the Swiss Federal Examination was developed but offers no empirical evaluation of the development process or the quality of the results.

None of the literature reviewed provided content validity evidence for the component parts of existing national licensing examinations. In other words, the literature does not help in establishing what should be tested in a national licensing examination – including the how and the why. This is in contrast to other tests, such as medical school examinations, language testing etc. where information is available.

Response process

Response process is concerned with how all the participants - candidates and officials - respond to the assessment. It is part of the quality control process.

Only three papers attempt to explore the validity of the response process.

Lillis, S., Stuart, M., Sidonie, Takai, N. (2012): 'New Zealand Registration Examination (NZREX Clinical): 6 years of experience as an Objective Structured Clinical Examination (OSCE)' describe the quality assurance process that validates this OSCE. The authors describe how this includes a full mock run through one week prior to the actual examination. This however is only achieved, in part, through the small numbers of candidates (28 in each cohort running 4 - 5 times a year) and that the examination takes place in one location in New Zealand.

Seyfarth et al., (2010): 'Grades on the Second Medical Licensing Examination in Germany Before and After the Licensing Reform of 2002.' aimed to statistically compare and assess the written and oral-practice grades of German students before and after licensing reform. The reform altered the format, scope, and timing of the administration of the medical licensing examinations. The first part of the examination and the written part of the second examination were removed. These were replaced by a written examination after the 'practical year' or pre-graduate internship. The second part of the examination was revised to include content from the clinical phase of training, after the first examination and including the internship. Using data from two German universities, the authors found the grades from the written exams did not differ in a statistically significant way (Seyfarth *et al.*, 2010). However a change in the clinical component grades had led to a *"significantly increased concordance between grades on the oral and written components of the examination."* They postulate first that the examiners in the oral-practical examination *post* revision might now expect more from the students because it had become a final examination. Second, candidates may have found it difficult to prepare for the new format.

Meanwhile, fears that the new clinical examination would lead to deterioration in the written examination scores were not confirmed.

Guttormsen, S., Beyeler, C., Bonvin, R., Feller, S., Schirlo, C., Schnabel, K., Schurter, T., Berendonk, C. (2013): 'The new licensing examination for human medicine: from concept to implementation.' The paper sets out and discusses procedures pertinent to the response process (e.g. candidate instructions, item scoring, station timings, rating scales, component weighting, assessor training, translation into multiple languages). The authors make some informal comparisons with similar examinations in the US and Canada.

Analysis

Two of the papers (Lillis et al., and Seyfarth et al.,) draw on limited data, which necessarily restricts the validity of the response process and the conclusions they reach. When the research was done, the revised German examination was clearly at an early stage of development and the authors acknowledge these limitations. They set out what would be required for better evidence. The paper on the Swiss Federal Licensing examination provides more extensive detail as the authors describe the development process.

Internal structure

Is the assessment structured in such a way as to make it reliable, reproducible, and generalizable? Are there any aspects of the assessment's structure that might induce bias?

Four papers in our review report the evidence to support the validity of internal structure.

Lillis, S., Stuart, M., Sidonie, Takai, N. (2012): 'New Zealand Registration Examination (NZREX Clinical): 6 years of experience as an Objective Structured Clinical Examination (OSCE).' The authors found the range of Cronbach alphas (Cronbach's alpha is a measure of test reliability) calculated over the prior 5 years were between 0.75 to 0.85. This means the internal consistency of the test is good. The authors also undertook a range of statistical analyses on the results of the examination as part of their quality control regimen. This involved the use of 'discrimination analysis' for each station: that is, does the design of the assessment discriminate against particular groups of students? At the end of each examination examiners and candidates complete anonymised feedback forms.

Ranney, R.R. (2006): 'What the Available Evidence on Clinical Licensure Exams Shows' identifies a number of studies in dentistry that indicate the low reliability of clinical licensure examinations. In relation to clinical licensure examinations the author draws on the findings of others and observes that many of the values and skills needed for safe practise are never tested. He notes that the unreliability of one-shot clinical examinations can often be traced to *"uncontrolled fluctuations in patients and circumstances of the examination"* (p149).

Harik, P., Clauser, B.E., Grabovsky, I., Margolis, M.J., Dillion, G.F., Boulet, J.(2006): 'Relationships among subcomponents of the USMLE Step 2 Clinical Skills examination, the

Step 1, and the Step 2 Clinical Knowledge examinations.' Harik et al. set out to examine the relationships between various sub-components of the USMLE among two candidate groups: first-time US medical students and first-time International medical graduates. They conclude from the statistical correlations that performance on Step 2 of the Clinical Skills examination, a simulation that uses a 'standardised patient' format to assess candidates interpersonal and communication skills:

> Is moderated by spoken English proficiency. This is consistent with expectations in that although this dimension is intended to be a separate and conceptually independent component of the test, for examinees with proficiency below a certain threshold it is unavoidable that English language skills will interfere with the ability to gather data, share information, and establish rapport. (Harik et al., 2006)

The authors also report on the statistical reliability of the subcomponents. The reliabilities for the subcomponents were acceptably high (>0.7) for overseas candidates, but two components were less reliable amongst home graduates. This latter fact they suggest provides a strong argument against combining the 'communication and interpersonal skills' and the 'spoken English proficiency' component scores for US medical graduates. In exploring the correlations between components of the Step 2 Clinical Skills and Step 2 Clinical Knowledge examinations Harik et al. found them to be positive but weak as the examinations are measuring two different things.

Guttormsen, S., Beyeler, C., Bonvin, R., Feller, S., Schirlo, C., Schnabel, K., Schurter, T., Berendonk, C. (2013): 'The new licensing examination for human medicine: from concept to implementation.' With regards to reliability in the Swiss Federal Licensing examination, the authors report the Cronbach's alpha as 0.91 for the written examination and 0.86-0.90 for the clinical skills examination. There was a moderate (0.52) correlation between the written and the clinical skills examinations. This verifies that they were measuring distinct competencies with some common ground.

Analysis

These four papers emphasise, in different ways, the rigour of current assessment processes and how educational assessors continue to re-evaluate their product and the constituent processes. Validity evidence is central to those efforts (Downing, 2003). Lillis (2012) does this by setting out the continuing quality control process for OSCEs. Harik et al., (2006) do likewise as they explore the statistical relationships between subcomponents of the USMLE Step 2. Ranney (2006), in contrast, reviews what was then (2006) the most up-to-date evidence on what makes one-shot, high stakes licensing examinations reliable. In so doing he concludes that in US dentistry at least, a reliable and valid examination has still to be devised. Guttormsen et al. (2013) provide good detail on all aspects of the process through which the examination was developed.

Relationship to other variables

The relationship to other variables is concerned with the connection between test scores and external variables. It seeks statistical, experimental, and observational or other evidence to confirm or deny any connections.

Ten papers explore the relationship between the results of licensing examinations and other measures of performance. These studies draw on empirical data from the PLAB, USMLE and the MRCP(UK).

Cuddy, et al., (2006): 'A Multilevel Analysis of the Relationships Between Selected Examinee Characteristics and United States Medical Licensing Examination Step 2 Clinical Knowledge Performance: Revisiting Old Findings and Asking New Questions.' The authors of this study examined the relationships between examinee characteristics and performance on the USMLE Step 2 Clinical Knowledge (CK) test. They used data from 54,487 examinees from 114 US accredited medical schools. Their results were consistent with previous examinee-level research, which found variations in Step 2 CK scores were associated with other variables such as the candidates' gender, Step 1 scores, time spent per item in the examination, the size of medical school, the mean Step 1 score, and the percentage of native English speakers. Women generally outperformed men on Step 2 CK.

Harik et al., (2006) 'Relationships among subcomponents of the USMLE Step 2 Clinical Skills examination, the Step 1, and the Step 2 Clinical Knowledge examinations.' The authors found that failure rates for international medical graduates were higher than for home graduates, and that this was partially attributable to poorer proficiency in spoken English.

Kenny et al., (2013): 'Associations between residency selection strategies and doctor performance: a meta-analysis.' The purpose of this study was to use meta-analysis to examine the relationships between a range of measures (including USMLE Step 1 & 2 scores) to assess applicants to residency programmes. The authors examine a variety of measures used to assess subsequent performance during residency and the doctor's subsequent career. They found scores in USMLE Steps 1 & 2 were significant and positively associated with in-training examinations, in-training evaluation reports, licensing examinations, and professional ratings. Associations with Step 1 & 2 scores were strongest for in-training and licensing examinations. Step 2 scores also showed an association with in-training evaluation reports, which was similar in strength to the association with licensing examinations.

Hecker K, & Violato, C. (2008): 'How much do differences in Medical Schools Influence Student Performance? A Longitudinal Study Employing Hierarchical Linear Modelling.' This study sought to determine whether students from different medical schools in the US, performed differently in Steps 1-3 of the USMLE over an eight-year period 1994-2004. The authors found the majority of the variation between medical schools in USMLE could be accounted for by student differences (85% of total variance), mostly MCAT scores (so examination performance prior to attending medical school). They also found that curriculum differences and school-level educational policies and educational innovations contributed only sporadically over the 8-year period. The authors noted a significant difference between schools when the geographic location and status (private/public) were taken into consideration.

McManus, I., & Wakeford, R. (2014): 'PLAB and UK graduates' performance on MRCP(UK) and MRCGP examinations: data linkage study' was a study to establish validity evidence for the Professional Language and linguistics Assessment Board (PLAB). The authors use correlation and multiple regression to assess whether the performance of IMGs who pass the Membership of the Royal Colleges of Physicians United Kingdom MRCP(UK) and Membership of the Royal College of General Practitioners MRCGP examinations is equivalent to UK graduates. The authors found PLAB scores correlated with MRCP(UK) and MRCGP, but that overall PLAB graduates' knowledge and skills at MRCP(UK) & MRCGP were poorer than UK graduates. Considerable increases in the PLAB pass marks would be needed to produce PLAB graduates of equivalent quality to UK graduates. The corollary of this is that it would reduce pass rates with subsequent *"implications for medical workforce planning."*

Ranney, R. (2006): 'What the Available Evidence on Clinical Licensure Exams Shows' summarises evidence, via a literature review, from a range of studies that examine the association between examination scores in dental licensure examinations and other assessments of clinical or factual knowledge. The results were mixed. Of 13 studies, 6 show positive associations, 2 show negative associations, and 5 show no association.

Stewart et al., (2005): 'Relationship Between Performance in Dental School and Performance on a Dental Licensure Examination: An Eight-Year Study' examined the association between academic performance in dental school and scores in the dental licensure examination. Using one-way ANOVAs to compare licensure examination scores and pass rates across quartile groups based on graduating GPA, they examined data relating to 524 graduates (1996-2003) from the University of Florida, College of Dentistry. The authors conclude that academic performance in dental school is predictive of licensing examination performance.

Guttormsen, S., Beyeler, C., Bonvin, R., Feller, S., Schirlo, C., Schnabel, K., Schurter, T., Berendonk, C. (2013): 'The new licensing examination for human medicine: from concept to implementation.' The researchers examined the pass rates for Swiss candidates and IMGs. They report these as 96.8-100% for Swiss candidates, 67.4% for IMGs in the written examination, 97.5-99.2% and 50% respectively in the clinical examination. IMGs mainly failed the clinical examination because of low scores on the history-taking, physical examination, and the diagnosis and management plan component rather than the communication skills component.

Tiffin et al., (2014): 'Annual Review of Competence Progression ARCP Performance of doctors who passed Professional and Linguistic Assessments Board (PLAB) tests compared with UK graduates' is an observational study using data relating to 53,436 UK based trainee doctors with at least one competency related ARCP outcome during the study period. Some of these trainees were IMGs who were registered having passed the PLAB test. The authors found that higher International English Language Test Scores (IELTS) and PLAB scores are predictive of better ARCP outcomes, with IMGs more likely to achieve poorer ARCP outcomes than UK graduates. They suggest that this disparity might be evened out by raising the pass marks for both parts of the PLAB test and raising the standards of English language competency. Another alternative is to devise and introduce a different test system.

Zahn et al., (2012): 'Correlation of National Board of Medical Examiners' Scores with the USMLE Step 1 and Step 2 Scores' explores the score data from 484 students graduating from 3 classes at the Uniformed Services University in 2008. The authors use statistical analysis to show a strong correlation between USMLE scores and NBME clerkship (clinical placement) scores. Most of the correlation is explained by performance in the primary care clerkship exam within a 2 year time period. The study confirms that students who do well in one test of knowledge are likely to do well in subsequent and similar tests of knowledge.

Analysis

A comparison of licensing examinations, including those specifically designed to assess IMGs, provides an opportunity to explore whether a national licensing examination brings unique or compelling validity evidence to the regulatory/safety debate.

The papers can be grouped into two areas of enquiry. First Hecker & Violato (2008), Ranney (2006), Stewart et al. (2005), Tiffin et al. (2014) and Zahn et al. (2012) all explore the relationship between medical school examination performance and established large scale testing i.e., USMLE. Overall they find, perhaps not surprisingly, that those who do well in examinations prior to and while at medical school also do well in later testing. Not all the difference in performance between students could be explained by previous examination performance difference though (Hecker & Violato, 2008). Kenny et al., (2013) provide similar evidence in their meta-analysis of USMLE performance and selection for residency programmes. Much of the validity evidence presented in these papers assures us that the specific assessments have validity in that they are able to identify candidates similarly to other similar tests.

Second, Cuddy et al., (2006), Harik et al., (2006), McManus & Wakeford (2014) and Tiffin et al. (2014) each demonstrate that the IMGs do less well in large scale testing. In assessments, such as the MRCP(UK), MRCGP and at the ARCP (Tiffin *et al.*, 2014), IMGs perform less well than UK graduates and this correlates with IMG performance on the PLAB. Both sets of authors argue that standards should be raised for IMGs by elevating the PLAB cut score or introducing different assessment methods.

Both papers demonstrate that the difference in performance scores between IMGs and UK graduates is not anomalous. The role that a national licensing examination might have therefore is in providing direct comparability between all doctors working in the UK. However, they also highlight the important consequences that might arise from attempting to raise standards in the ways they suggest, as some IMGs may not wish or be able to work in the UK thereby leading to workforce shortages.

Cuddy et al., (2006) and Harik et al., (2006) provide some similar evidence from their analyses of the USMLE. Once again the effect that a lack of proficiency in spoken English has is evident. However, as Cuddy et al., (2006) observe, other factors such as gender and time spent per item in the examination also have some effect. As with the other papers, these two only identify potential statistical links between these particular variables. In contrast, Guttormsen et al., (2013) identified that IMGs did less well than Swiss candidates in the Federal Licensing Examination, but that the low scores for IMGs were in areas other than the communication skills component.

Consequences

Consequences or evidence of impact is concerned with the intended or unintended consequences assessment may have on participants or wider society. It may include whether assessments provide tangible benefits or whether they have an adverse or undesirable impact.

Fifteen papers discuss the consequential validity or impact of licensing examinations.

Ahn, D., & Ahn, S. (2007): Reconsidering the Cut Score of the Korean National Medical Licensing Examination.' The authors argue the cut score in the Korean National Medical Licensing Examination was arbitrarily set at 60% during Japanese colonial rule. They draw on validity and standard setting evidence from elsewhere. After surveying Korean psychometricians, medical educators, and examiners for their views, the authors argue the Bookmark and modified Angoff standard-setting approaches offer more useful alternatives to setting cut scores. They conclude with a discussion about the feasibility challenges of undertaking complex standard setting on large scale examinations.

Green, M., Jones, P., Thomas Jr, J.X. (2009): 'Selection Criteria for Residency: Results of a National Program Directors Survey.' This study reports on the results of an email and postal survey completed by National Program directors. The purpose of the study was to assess the perceptions of programme directors as to the relative importance of various criteria in the selection process, including USMLE Step 1 & 2 scores. 2,528 programme directors were sent a survey (85% of the 2,980 listed) in 21 selected specialities. The authors conclude that USMLE Step 1 & 2 scores are regarded as highly important criteria in selecting medical students for postgraduate training. USMLE Step 1 & 2 scores were significantly higher in 'most competitive' specialities. Thus, higher scores on USMLE Steps 1 & 2 are positively associated with the likelihood of gaining a residency programme place in those specialties.

Holtzman et al., (2014): 'International variation in performance by clinical discipline and task on the United States Medical Licensing Examination Step 2 Clinical Knowledge Component.' uses descriptive statistics to examine variations in the USMLE Step 2 clinical knowledge examination between US graduates and IMGs from various countries between 2008 & 2010. They found that IMGs' perform less well than US graduates. They postulated that the poorer performance of IMGs may arise from differences in curricula, clinical experiences, and the patient populations encountered by trainees. Other reasons suggested are: cultural differences, differential effects of English as a second language, structure and quality of educational programmes, and differences in how medical schools prepare students for the three step USMLE. No evidence is offered to back these possible reasons for the disparity in performance.

Kenny et al., (2013): 'Associations between residency selection strategies and doctor performance: a meta-analysis.' The paper provides some indirect evidence for consequences. Many of the studies in the meta-analysis use USMLE scores as part of the selection measures used in the ranking process for residency programmes. This is indicative of the widespread use of these scores in selections processes. Thus, performance in the USMLE can have consequences for a doctor's future career.

Kugler, A.D, & Sauer R.M. (2005): 'Doctors without Borders? Relicensing Requirements and Negative Selection in the Market for Physicians' considers national licensing from an economic perspective. The authors use official statistics from Israel on doctors arriving in the country from the former USSR. Depending upon length of previous medical experience, immigrant doctors seeking a licence to practise had to (a) take an exam, or (b) work under supervision for six months. The authors use this data to develop a model of optimal licence acquisition. They found that 73% of the less experienced doctors obtained a licence through the examination route. 89% of the more experienced doctors were assigned to the supervision route. The policy implications of the study are that:

> *… lowering the costs to immigrant physicians of acquiring a medical licence may raise average physician quality … assignment to the observation track has more of an*

> *impact on the probability of licence acquisition than on the probability of physician employment.* (457)

The authors conclude the economic benefits of obtaining a licence were generally high, but earnings in unlicensed occupations were better for those who did not obtain a licence than those who did. A consequence of this is that it may induce more broadly skilled doctors to seek unlicensed occupations.

Lillis, S., Stuart, M., Sidonie, Takai, N. (2012): 'New Zealand Registration Examination (NZREX Clinical): 6 years of experience as an Objective Structured Clinical Examination (OSCE)' utilise a combination of a Borderline groups method for their dynamic (interactive with an examiner present) OSCE stations and a modified Angoff method for the static (slide or image based) OSCE stations. The cut score is adjusted for the Standard Error of Measurement to allow for any uncertainty of scores. All their stations have equal weight and there are no 'killer stations'. However, a 'critical incident' policy was introduced for instances where there has been a clear breach of expected professional standards.

Margolis et al., (2010): 'Validity Evidence for USMLE Examination Cut Scores: Results of a Large Scale Survey' used a large scale questionnaire 1,500 stakeholders across medical training bodies in the US on the cut score (pass mark) of USMLE Steps 1-3. The survey produced a low response from examinees and a good response from authorities. The results were mapped to Kane's measures of validity. Responders felt failure rates were about right for the exam (6-7% Step 1, 4-6% Step 2, 4-5% Step 3). Some thought that because nearly all candidates ultimately pass (<1% after n retakes) the cut score might be too low. The authors also found that residency programme directors (those charged with overseeing doctors once in practice) wanted to see a higher failure rate.

Musoke, S. (2012): 'Foreign Doctors and the Road to a Swedish Medical License' arises from a Bachelor thesis in Global Development that contains empirical qualitative data in the form of recorded interviews with five non-European doctors who were trying to obtain a Swedish medical licence. The thesis also draws on qualitative data from a seminar with Swedish doctors about the process that foreign doctors must go through to work in Sweden.

The thesis contains verbatim quotes from the five non-European doctors. The similarity of experience among the participants adds credibility to the data and their observations. The five non-European doctors in the study felt disadvantaged or disfavoured by the Swedish licensure process. European doctors, in comparison, were felt to be favoured by the system. The participants stopped short of saying they felt discriminated against. The researcher concludes the system is flawed, confusing, frustrating, and overly long. Some cross referencing with studies in other countries gives additional validity to the conclusions drawn.

Norcini et al., (2014): 'The relationship between licensing examination performance and the outcomes of care by international medical school graduates' is a US study focused on the performance of IMGs in the USMLE Step 2 Clinical Knowledge examination and whether there was any *"relationship between the scores on the Step 2 CK examination and in-hospital mortality for patients with CHF [chronic heart failure] or AMI [acute myocardial infarction]"* (p1157). This retrospective observational study uses descriptive statistics and a multivariate analysis which found that each additional point on the examination was associated with a 0.2% decrease in mortality. The size of the effect was noteworthy, with each standard deviation (roughly 20 points) equivalent to a 4% change in mortality risk. The authors conclude that the findings *"… provide evidence for the validity of the Step 2 CK scores … the results support the use of the examination as an effective screening strategy for licensure"* (p1157). The authors acknowledge the limitations of the research data and that other factors might also explain these results. They suggest further research is required as there may be other factors acting to explain the difference in patient outcomes.

Ranney, R. (2006): 'What the Available Evidence on Clinical Licensure Exams Shows' draws on literature and the results of a survey of Dental School Deans. He concludes that dental licensure examinations in the US and Canada lack the necessary reliability and validity required for 'one-off', high states examinations. This, he suggests, has consequences for those taking the examinations and for those with the mandate to ensure patient safety and professional competence. Echoing the view of Dental School Deans who *"thought it was important to realize change in licensure processes for Dentists"* (p152), he recommends a reliable and valid licensure examination is developed.

Stewart et al., (2005): 'Relationship Between Performance in Dental School and Performance on a Dental Licensure Examination: An Eight-Year Study' identify that the weighting of examination components, and variation in pass rates between these components are influential on student outcomes. The authors suggest Dental Colleges take these findings into account when preparing students for the dental licensure examinations.

Sutherland, K., & Leatherman, S. (2006): 'Regulation and Quality Improvement A Review of the Evidence' draws on a systematic review (including grey literature) on regulatory interventions in healthcare systems across the world. The purpose of the study was to determine 'what works.' The authors group the literature under three headings: 'Institutional regulation', 'Professional regulation', and 'Market regulation.' They conclude there is little evidence to answer the question of 'what works?' With regards to physician licensure, the authors observe, *"… there is little evidence available about its impact on quality of care"* (p8).

Guttormsen, S., Beyeler, C., Bonvin, R., Feller, S., Schirlo, C., Schnabel, K., Schurter, T., Berendonk, C. (2013): 'The new licensing examination for human medicine: from concept to implementation.' The authors describe the format of candidate feedback on performance. Standard setting for the examination used both Angoff and Hofstee methods, but how these were combined is not clear. Standard setting for the clinical examinations used borderline regression.

Tamblyn et el., (2007): 'Physician Scores on a National Clinical Skills Examination as Predictors of Complaints to Medical Regulatory Authorities' report on a longitudinal study to assess whether patient-physician communication exam scores of candidates who passed the Medical Council of Canada (MCC) clinical skills exam from 1993 to 1996, could predict future complaints in later medical practice. The physician cohort comprised 3,424 physicians. Those with lower scores in the MCC clinical exam (the bottom 2.5%) are more likely to have complaints made against them in future practice. Most complaints arose through 'communication problems.' The complaint rate observed was 0.0491 per physician. The authors conclude, *"Scores achieved in patient-physician communication and clinical decision making on a national licensing examination predicted complaints to medical regulatory authorities."* The correlation between communication skills and complaints study demonstrates the need to test communication skills. It does not establish causation however. The authors acknowledge the limitations of the *"poor-to-moderate reliability of the communication score component of the examination …"* and how the use of *"practice-years as a denominator for estimating the rate of complaints would not take into account the frequency of patient contact, the type of patients, and the procedures performed …"*

Wenghofer et al., (2009): 'Doctors Scores on National Qualifying Examinations Predict Quality of Care in Future Practice' is a Canadian study to determine whether national licensing examinations (the MCCQE Parts 1 & 2) predict the quality of care delivered by doctors in their future practice. The authors use multivariate logistic regression on data from a cohort of doctors. The findings suggest doctors in the bottom quartile of each examination are more likely to be assessed as providing an *"unacceptable quality of care assessment."* Although the authors acknowledge that, *"relatively few quality of care assessments resulted in unacceptable outcomes in the study population, which resulted in wide confidence intervals around the estimates of the examination and peer assessment relationship"* overall, they insist *"Doctors' scores on MCCQE1 are significant predictors of quality-of-care problems based on regulatory, practice-based peer assessment."* They also acknowledge there are likely to be, *"additional covariate factors not included in our model that may influence the relationship between qualifying examination scores and practice performance …"*

Analysis

Sutherland & Leatherman concluded in 2006 that "*there is little evidence available about [national licensing examinations] impact on quality of care"* across the international healthcare systems (Sutherland, 2006). Since then researchers have tried to make these links. Both Norcini et al. (2014) and Tamblyn et al. (2007) explored the correlation between performance on national licensing examinations (in the US and Canada respectively) and subsequent specific patient outcomes (Norcini *et al.*, 2014) or rates of complaints (Tamblyn *et al.*, 2007). What they found are excellent arguments for the importance of medical education and testing, however their findings are limited to establishing correlations between testing and outcomes. The papers are though important and contribute to the content validity needs of any examination process. They demonstrate the need for communication as well as knowledge testing.

The papers by (Green, Jones & Thomas, 2009) and (Kenny, McInnes & Singh, 2013) demonstrate the career consequences that come from how well candidates pass the USMLE. This is a confounder in understanding the impact of examinations on patient outcomes. Those who score higher in the USMLE end up in the better healthcare institutions – these institutions are likely to play as big a part in patient outcomes as the individuals employed there.

It is interesting to note that, other than Stewart et al's (2005) suggestion that Dental Colleges look at the dental examination in Florida when preparing their students, there appears to be no empirical evidence as to the impact of licensing examinations on prior education programmes.

There are an important group of papers that carry forward the concerns around balancing the idea of protecting the public whilst fulfilling workforce needs. Margolis et al. (2010) identifies concerns that nearly everyone who takes USMLE passes it in the end. While authors continue to find that IMG doctors do less well there remains a lack of evidence as to why this is.

Musoke (2012) raises the issue of IMGs being stigmatised and disadvantaged as they negotiate a confusing bureaucratic process. It may be that national licensing examinations might not lead to equality. Kugler & Sauer (2005) argue the economic case that IMG doctors simply find ways to work around the system or seek alternative employment if additional barriers are placed before them. An unintended consequence of a UK national licensing examination under current EU law might be that IMGs seek citizenship in an EU partnership country and then enter the UK thereby bypassing any new UK licensing examination requirements.

Finally, (Guttormsen *et al.*, 2013) in setting out the evolution of the Swiss Federal Licensing Examination provide some European data. Their observation that IMGs did less well than Swiss candidates fits with patterns found elsewhere.

4. Discussion

The literature collected during the review is diverse in its quality, its methodology, and the evidence it provides for the validity of medical licensure examinations. It also offers a number of perspectives on the impact of licensure examinations. What is clear from the literature we garnered, is that the testing and assessment of licensure examinations, especially the North American USMLE model, is now a sophisticated enterprise. For this reason perhaps, the technical aspects are reasonably well evidenced (Bajammal *et al.*, 2008; CUP, 2008; Sonderen *et al.*, 2009). From a pedagogic and a legal standpoint this makes them defensible. The industrial scale of licensure examinations on the North American continent means a large amount of statistical data is available for analysis. Consequently, there is a tendency for the literature to explore the North American experience.

In contrast, some of the broader and bigger claims made for licensure examinations are less well evidenced. In particular those made for greater patient safety (McMahon & Tallia, 2010; Melnick, 2009; Norcini *et al.*, 2014), improved quality of care (Wenghofer *et al.*, 2009), and identification of doctors likely to face disciplinary action (Tamblyn *et al.*, 2007). While there is no denying that a statistical correlation appears to exist between a candidate's performance in a national licensing examination and some aspects of future practice, the large number of variables unaccounted for by this research limits their interpretation. For example, as the studies by (Green, Jones & Thomas, 2009) and (Kenny, McInnes & Singh, 2013) demonstrate, candidates with lower scores tend to work in less respected institutions. Lower scores can result in graduates working in less desirable or poorer performing organisations (Noble, 2008). Furthermore, in spite of the claims made by those studies that claim patient care and poor disciplinary records can be predicted from pass scores, a comprehensive review by Sutherland & Leatherman (2006) on whether regulation improves healthcare found 'sparse' evidence to back such claims. Our review supports that conclusion.

In unpacking the literature we found lively debate on licensure in many healthcare professions – particularly in the US. In US dentistry for example a fractious debate around licensure examinations, spurred in part by legislative and regulatory 'turf wars', has been taking place for some time. The circularity of that debate, which mirrors the circularity of the current debate among doctors, resulted in some US dental bodies breaking the argumentative impasse by legislating for an alternative to a national dental licensure examination – the 'residency pathway to licensure.' Within a US context, this was felt to be a 'sea-change' in regulatory thinking (Ferris, 2006).

For those who argue against national licensure examinations (Harden, 2009) or those who hedge on the topic (Schuwirth, 2007; van der Vleuten, 2013; van der Vleuten, 2009) a similar problem with regards to evidence arises. These well-informed academics draw on a variety of research studies to either rebut the pro-licensure lobby or evaluate the pros and cons of national licensure through the use of a North American style licensing examination. But, as they themselves point out, *unequivocal* evidence is lacking and a knowledge gap identified (Boulet & van Zanten, 2014).

The review also suggests a significant knowledge gap exists around the impact of licensure examinations on IMGs. Whilst a strong body of statistical evidence exists to show IMGs perform less well in licensure examinations than candidates from the host countries (Guttormsen *et al.*, 2013; Holtzman *et al.*, 2014), the reasons for this phenomenon remain unclear. In view of the significant part IMGs play in the physician workforce of many countries including the UK, *and* the apparent difficulties they present to regulators, this is an area of research that needs to be better understood.

What research there is (at least that meet our inclusion criteria) suggests IMGs (Sonderen *et al.*, 2009) and migrant physicians (Kugler A.D & Sauer, 2005) may, for a number of reasons, work in occupations that do not necessarily match their skills or qualifications. If this is so, and if licensure examinations are a contributory factor, then in a world where physician shortages exist it seems appropriate to explore this further.

Of course such issues raise difficult questions about inclusion, exclusion, and fairness (McGrath, Wong & Holewa, 2011). Musoke's (2012) research on the experiences of IMGs in Sweden indicates that the regulatory regime in force there (which is not dissimilar to regulatory processes across Europe and elsewhere) may actively disadvantage competent practitioners – even those who are competent Swedish speakers. She, and those she researched, viewed the Swedish system as flawed, overlong, and frustrating. Other research indicates this is not a just a Swedish problem (Kovacs *et al.*, 2014).

In the same vein, several Canadian studies in the review outline similar difficulties for IMGs, and those who employ them, in negotiating licensure examinations (Audas, 2005; Maudsley, 2008). These studies provide some descriptive evidence for the way in which practitioners, provincial licensing authorities, and employers actually use the system to balance the demands that arise from physician shortages. Meanwhile, McGrath, Wong et al's (2011) assessment of Canadian and Australian approaches to IMGs reveals some fundamental ideological differences in how IMGs fit into the workplace landscape of each country. In Canada the approach is one of assimilation. In Australia regulators foster a parallel but separate workforce culture. The importance in this distinction is that it may affect where an increasingly mobile workforce may choose to migrate to.

Overall, our review concludes that the debate around licensure examinations is strong on opinion but weak on validity evidence. This is especially true of the wider claims that licensure examinations improve patient safety and practitioner competence. What is clear is that where national licensing and other large scale examinations exist there is a relationship between examination performance in those examinations and similar subsequent ones. There is also a less well-explored relationship between examination performance and some patient outcomes and rates of complaints. Nowhere, as yet, has staged a national licensing examination to establish in the style of a randomised control trial whether such an examination impacts upon measures of interest such as patient outcomes. Until more evidence for these aspects of the licensing examination is produced, the debate will continue, be prolonged and circular.

How might an evidence base for national licensing examinations be better established? There is no doubt that any new examination would need to be developed in line with good assessment principles. The APA validity framework provides an internationally recognised approach to establishing validity evidence, please also see Shaw et al. (Shaw S, Crisp V & Johnson N, 2012).

But in order to understand any new initiative and collect validity evidence to assure the regulator, the public and the profession there are three main approaches:

1. There could be the establishment of a basic *process* evaluative framework that seeks to understand the outputs of the new examination. This would include much of the evidence as highlighted in this literature review. Importantly this would mostly be retrospective and seek traditional validity evidence.
2. Any new initiative could be developed in conjunction with an *outcomes* evaluative framework and not *post*-hoc (as in [1]). This would require selecting measures before and after any intervention – in this case a national licensing examination – to see if the measure changes as a result.
3. Lastly, there might be opportunities to explore the use of trialist methodologies, such as randomised control trial designs, to establish whether a new national licensing examination really produces added value in terms of patient care and outcomes. While this might initially be seen as difficult the opportunity to establish a control group so that comparison can be made between those that go through a national licensing examination and those who do not is central to taking the arguments forward. Control studies can include step-wedged designs (sub-groups going through the intervention at different points in time) and cross-over trials (where the control group undertake the intervention at the end of the trial) so that ultimately all subjects have experienced the intervention.

Within each of these approaches, and in line with the GMC's commitment to exploring differential attainment, there are opportunities to specifically determine *why* IMGs do less well in national licensure examinations. Data would be generated that would allow all home students as well as IMGs to be compared on performance in relation to variables such as ethnicity and gender.

Ultimately to understand fully this would require a mixed-methods approach to explore not simply statistical differences but the underlying causes. It should form part of any evaluative framework. This could be important in ensuring that licensure examinations do not act as an inappropriate barrier to IMG entry into the physician workforce and to learn more about the real-world impact of licensure processes on IMGs.

6. References

Ahn, D. S. & Ahn, S. (2007) 'Reconsidering the cut score of Korean National Medical Licensing Examination'. *Journal Of Educational Evaluation For Health Professions*, 4 pp 1-1.

Audas, R., Ross, A, Vardy, D. (2005) 'The use of provisonally licensed international medical graduates in Canada'. *Canadian Medical Association Journal*, 173 pp 1315 - 1316.

Avery, M. D., Germano, E. & Camune, B. (2010) 'Midwifery Practice and Nursing Regulation: Licensure, Accreditation, Certification, and Education'. *Journal of Midwifery & Women's Health*, 55 (5). pp 411-414.

Bajammal, S., Zaini, R., Abuznadah, W., Al-Rukban, M., Aly, S., Boker, A., Al-Zalabani, A., Al-Omran, M., Al-Habib, A., Al-Sheikh, M., Al-Sultan, M., Fida, N., Alzahrani, K., Hamad, B., Al Shehri, M., Abdulrahman, K., Al-Damegh, S., Al-Nozha, M. & Donnon, T. (2008) 'The need for national medical licensing examination in Saudi Arabia'. *BMC Medical Education*, 8 (1). pp 53.

Bettany-Saltikov, J. (2010) 'Learning how to undertake a systematic review: part 1'. *Nursing Standard*, 24 (50). pp 47-56.

Borow, M., Levi, B. & Glekin, M. (2013) 'Regulatory tasks of national medical associations - international comparison and the Israeli case'. *Israel Journal of Health Policy Research*, 2 (1). pp 8.

Boulet, J. & van Zanten, M. (2014) 'Ensuring high-quality patient care: the role of accreditation, licensure, specialty certification and revalidation in medicine'. *Medical Education*, 48 (1). pp 75-86.

Calnon, W. R. (2006) 'The Residency Pathway to Dental Licensure: The Paradigm Shift from Inception to Policy'. *Journal of Evidence Based Dental Practice*, 6 (1). pp 138-142.

Chenot, J.-F. (2009) 'Undergraduate medical education in Germany'. *GMS German Medical Science*, 7

Cooper, S. L. (2005) 'The licensure mobility experience within the United States'. *Optometry - Journal of the American Optometric Association*, 76 (6). pp 347-352.

Cosby Jr, J. C. (2006) 'The American Board of Dental Examiners Clinical Dental Licensure Examination: A Strategy for Evidence-Based Testing'. *Journal of Evidence Based Dental Practice*, 6 (1). pp 130-137.

CUP, U. (2008) 'Comprehensive Review of USMLE Summary of Final Report and Recommendations'.

de Vries, H., Sanderson, P., Janta, B., Rabinovich, L., Archontakis, F., Ismail, S., Klautzer, L., Marjanovic, S., Patruni, B., Puri, S., Tiessen, J. (2009) 'International Comparison of Ten Medical Regulatory Systems'.

Downing, S. M. (2003) 'Validity: on the meaningful interpretation of assessment data'. *Medical Education*, 37 (9). pp 830-837.

Doyle, S. (2010) 'One-stop shopping for international medical graduates'. *Canadian Medical Association Journal*, 182 (15). pp 1608.

Ferris, R. T. (2006) 'A Sea-Change in American Dentistry: A National Clinical Licensure Examination, or a Residency-Based Pathway?'. *Journal of Evidence Based Dental Practice*, 6 (1). pp 129.

Gorsira, M. (2009) 'The utility of (European) licensing examinations. AMEE Symposium, Prague 2008'. *Medical Teacher*, 31 (3). pp 221-222.

Grant, M. J. & Booth, A. (2009) 'A typology of reviews: an analysis of 14 review types and associated methodologies'. *Health Information & Libraries Journal*, 26 (2). pp 91-108.

Green, M., Jones, P. & Thomas, J. X. J. (2009) 'Selection Criteria for Residency: Results of a National Program Directors Survey'. *Academic Medicine*, 84 (3). pp 362-367.

Guttormsen, S., Beyeler, C., Bonvin, R., Feller, S., Schirlo, C., Schnabel, K., Schurter, T. & Berendonk, C. (2013) 'The new licencing examination for human medicine: from concept to implementation'. *Swiss Med Wkly*, 143 (w13897). pp 1-10.

Harden, R. M. (2009) 'Five myths and the case against a European or national licensing examination'. *Medical Teacher*, 31 (3). pp 217-220.

Harik, P., Clauser, B. E., Grabovsky, I., Margolis, M. J., Dillon, G. F. & Boulet, J. R. (2006) 'Relationships among Subcomponents of the USMLE Step 2 Clinical Skills Examination, The

Step 1, and the Step 2 Clinical Knowledge Examinations'. *Academic Medicine*, 81 (10). pp S21-S24.

Hecker, K. & Violato, C. (2008) 'How Much Do Differences in Medical Schools Influence Student Performance? A Longitudinal Study Employing Hierarchical Linear Modeling'. *Teaching and Learning in Medicine*, 20 (2). pp 104-113.

Holtzman, K. Z., Swanson, D. B., Ouyang, W., Dillon, G. F. & Boulet, J. R. (2014) 'International Variation in Performance by Clinical Discipline and Task on the United States Medical Licensing Examination Step 2 Clinical Knowledge Component'. *Academic Medicine*, Publish Ahead of Print pp 10.1097/ACM.0000000000000488.

Kane, M., Crooks, T. & Cohen, A. (1999) 'Validating Measures of Performance'. *Educational Measurement: Issues and Practice*, 18 (2). pp 5-17.

Kenny, S., McInnes, M. & Singh, V. (2013) 'Associations between residency selection strategies and doctor performance: a meta-analysis'. *Medical education*, 47 (8). pp 790-800.

Kovacs, E., Schmidt, A. E., Szocska, G., Busse, R., McKee, M. & Legido-Quigley, H. (2014) 'Licensing procedures and registration of medical doctors in the European Union'. *Clinical Medicine, Journal of the Royal College of Physicians of London*, 14 (3). pp 229-238.

Kugler A.D & Sauer, R. M. (2005) 'Doctors without Borders? Relicensing Requirements and Negative Selection in the Market for Physicians'. *Journal of Labor Economics*, 23 (3). pp 437-465.

Lee, Y. S. (2008) 'OSCE for the Medical Licensing Examination in Korea'. *Kaohsiung Journal of Medical Sciences*, 24 (12). pp 646-650.

Lehman, E. P. & Guercio, J. R. (2013) 'The Step 2 Clinical Skills Exam — A Poor Value Proposition'. *New England Journal of Medicine*, 368 (10). pp 889-891.

Leitch, S. & Dovey, S. M. (2010) 'Review of registration requirements for new part-time doctors in New Zealand, Australia, the United Kingdom, Ireland and Canada'. *Journal of primary health care*, 2 (4). pp 273-280.

Lillis, S., Stuart, M., Sidonie, Takai, N. (2012) 'New Zealand Registration Examination (NZREX Clinical): 6 years of experience as an Objective Structured Clinical Examination (OSCE)'. *The New Zealand Medical Journal*, 125 (1361). pp 74 - 80.

Lopez-Valcarcel, B. G., Ortún, V., Barber, P., Harris, J. E. & García, B. (2013) 'Ranking Spain's Medical Schools by their performance in the national residency examination'. *Revista Clínica Española*, 213 (9). pp 428-434.

Margolis, M. J., Clauser, B. E., Winward, M. & Dillon, G. F. (2010) *Validity Evidence for USMLE Examination Cut Scores: Results of a Large-Scale Survey. [Miscellaneous Article].* Academic Medicine October 2010;85(10) Supplement, RIME: Proceedings of the Forty-Ninth Annual Conference November 7-November 10, 2010:S93-S97.

Maudsley, R. F. (2008) 'Assessment of International Medical Graduates and Their Integration into Family Practice: The Clinician Assessment for Practice Program'. *Academic Medicine*, 83 (3). pp 309-315 310.1097/ACM.1090b1013e318163710f.

McGrath, P., Wong, A. & Holewa, H. (2011) 'Canadian and Australian licensing policies for international medical graduates: a web-based comparison'. *Education for health (Abingdon, England)*, 24 (1). pp 452.

McMahon, G. T. & Tallia, A. F. (2010) 'Perspective: Anticipating the challenges of reforming the United States medical licensing examination'. *Academic Medicine*, 85 (3). pp 453-456.

McManus, I. C. & Wakeford, R. (2014) 'PLAB and UK graduates' performance on MRCP(UK) and MRCGP examinations: data linkage study'. *BMJ*, 348 pp g2621.

Melnick, D. E. (2009) 'Licensing examinations in North America: Is external audit valuable?'. *Medical Teacher*, 31 (3). pp 212-214.

Messick, S. (1995) 'Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning'. *American Psychologist*, 50 (9). pp 741-749.

Musoke, S. B. (2012) 'Foreign Doctors and the Road to a Swedish Medical License Experienced barriers of doctors from non-EU countries'.

Neilson, R. (2008) *Authors have missed gap between theory and reality.* http://www.bmj.com/content/337/bmj.a1783.full?sid=f010428c-6134-4bd3-bbbd-31966df6c19b edn. vol. 337.

Neumann, L. M. & Macneil, R. L. (2007) 'Revisiting the National Board Dental Examination'. *Journal of dental education*, 71 (10). pp 1281-1292.

Noble, I. S. G. (2008) *Are national qualifying examinations a fair way to rank medical students? No.* vol. 337.

Norcini, J. J., Boulet, J. R., Opalek, A. & Dauphinee, W. D. (2014) 'The Relationship Between Licensing Examination Performance and the Outcomes of Care by International Medical School Graduates'. *Academic Medicine*, 89 (8). pp 1157-1162 1110.1097/ACM.0000000000000310.

Nowakowski, M. (2013) 'National Medical License Exams in Poland'. *Jahrestagung der Gesellschaft für Medizinische Ausbildung (GMA)*,

Pavão Martins, I. (2013) *Admission to Residence Training in Portugal: Analysis of the National Exam Results between 2006 and 2011.* 2013. vol. 26.

Philipsen, N. C. & Haynes, D. (2007) 'The Multi-State Nursing Licensure Compact: Making Nurses Mobile'. *The Journal for Nurse Practitioners*, 3 (1). pp 36-40.

Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., Britten, N. Roen, K., Duffy, S. (2006) 'Guidance on the Conduct of Narrative Synthesis in Systematic Reviews. A Product from the ESRC Methods Programme'.

Ranney, R. R. (2006) 'What the Available Evidence on Clinical Licensure Exams Shows'. *Journal of Evidence Based Dental Practice*, 6 (1). pp 148-154.

Rehm, L. P. & DeMers, S. T. (2006) 'Licensure'. *Clinical Psychology: Science and Practice*, 13 (3). pp 249-253.

Ricketts, C. & Archer, J. (2008) *Are national qualifying examinations a fair way to rank medical students? Yes.* vol. 337.

Rowe, A. G.-B., M. (2005) 'Regulation and licensing of Physicians in the WHO European Region'.

Schuwirth, L. (2007) 'The need for national licensing examinations'. *Medical Education*, 41 (11). pp 1022-1023.

Seyfarth, M., Reincke, M., Seyfarth, J., Ring, J. & Fischer, M. R. (2010) 'Grades on the Second Medical Licensing Examination in Germany Before and After the Licensing Reform of 2002: A study in Two Medical Schools in Bavaria'. *Dtsch Arztebl International*, 107 (28-29). pp 500-504.

Shaw S, Crisp V & Johnson N (2012) 'A framework for evidencing assessment validity in large-scale, high-stakes international examinations'. *Assessment in Education: Principles, Policy & Practice*, 19 (2). pp 159-176.

Sonderen, M. J., Denessen, E., Cate, O. T. J. T., Splinter, T. A. W. & Postma, C. T. (2009) 'The clinical skills assessment for international medical graduates in the Netherlands'. *Medical Teacher*, 31 (11). pp e533-e538.

Stewart, C. M., Bates, R. E. & Smith, G. E. (2005) 'Relationship Between Performance in Dental School and Performance on a Dental Licensure Examination: An Eight-Year Study'. *Journal of dental education*, 69 (8). pp 864-869.

Sutherland, K. L., S. (2006) 'Regulation and Quality Improvement a review of the evidence'. *Health Foundation*,

Tamblyn, R., Abrahamowicz, M., Dauphinee, D. & et al. (2007) 'PHysician scores on a national clinical skills examination as predictors of complaints to medical regulatory authorities'. *JAMA*, 298 (9). pp 993-1001.

Tiffin, P. A., Illing, J., Kasim, A. S. & McLachlan, J. C. (2014) 'Annual Review of Competence Progression (ARCP) performance of doctors who passed Professional and Linguistic Assessments Board (PLAB) tests compared with UK medical graduates: national data linkage study'. *BMJ*, 348 pp g2622.

UNDP (2014) 'Human Development Report 2014 Sustaining Human Progress: Reducing Vulnerabilities and Building Resilience '.

van der Vleuten, C. (2013) *National licensing examinations and their challenges.* vol. 1.

van der Vleuten, C. P. M. (2009) 'National, European licensing examinations or none at all?'. *Medical Teacher*, 31 (3). pp 189-191.

Waldman, H. B. & Truhlar, M. R. (2013) 'Impact of residency requirement for dental licensure: an update'. *The New York state dental journal*, 79 (5). pp 30-32.

Wenghofer, E., Klass, D., Abrahamowicz, M., Dauphinee, D., Jacques, A., Smee, S., Blackmore, D., Winslade, N., Reidel, K., Bartman, I. & Tamblyn, R. (2009) 'Doctor scores on national qualifying examinations predict quality of care in future practice'. *Medical Education*, 43 (12). pp 1166-1173.