Clinical Medicine 2010, Vol 10, No 2: 141–44

A critique of the specialty certificate examinations of the Federation of Royal Colleges of Physicians of the UK

John Cookson

ABSTRACT – The Federation of Royal Colleges of Physicians of the UK has developed a programme to deliver specialty certificate examinations. These are knowledge-based examinations to be passed by all senior trainees in most medical specialties seeking a Certificate of Completion of Training (CCT). These examinations have been evaluated on their validity, reliability, educational impact, cost effectiveness, acceptability and standard setting methodology on the basis of internal evidence and the results of a published pilot. The evidence so far suggests that though reasonable reliability (reproducibility) can be achieved, validity (testing what is intended) may be lacking. Educational impact, cost effectiveness, and acceptability require more evidence. Consistency in standard setting is difficult.

KEY WORDS: assessment, postgraduate, reliability, validity

Introduction

The Federation of Royal Colleges of Physicians of the UK, in association with the specialist societies, has developed a programme to deliver specialty certificate examinations. The examination is summarised as:

> *A specialty certificate examination is now a compulsory component of assessment for Certificate of Completion of Training (CCT) for all UK trainees whose specialist training began in or after August 2007 and is in one of the following specialties: acute medicine; clinical pharmacology and therapeutics; dermatology; endocrinology and diabetes; gastroenterology, geriatric medicine; infectious diseases; medical oncology; nephrology (renal medicine); neurology; respiratory medicine and rheumatology.*[1]

Its purpose is to:

- identify practising trainees who have acquired the necessary professional knowledge and problem-solving skills to enable them to practise independently as specialists
- evaluate the professional competence of medical graduates during specialist training in areas such as clinical science, epidemiology and statistics.[1]

The specialty certificate examinations meet requirements for knowledge-based assessments that are a part of the curricula submitted to the Postgraduate Medical Education and Training Board (PMETB). UK trainees who have completed MRCP(UK) would normally take the specialty certificate examination during higher specialist training, and should have made at least one attempt by the time of their penultimate year assessment. It forms part of a suite of assessment tools which include clinical assessments (mini clinical evaluation exercise (miniCEX) and direct observation of procedural skills (DOPS)), but remains independent of the Record of In-training Assessment (RITA).

This paper offers a critique of the specialty certificate examinations. It is largely based on the information publically available on the websites of the Royal College of Physicians (RCP) and others. These include the Joint Committee on Higher Medical Training report of the pilot project,[2] details of the curricula from the Joint Royal College of Physicians Training Board[3] and details of the specialty specific examinations.[4]

Characteristics of a good assessment system

Schuwirth and van der Vleuten have proposed a number of characteristics of a good assessment system which have been widely accepted.[5] These are validity, reliability, educational impact, cost effectiveness and acceptability. These are very similar to those of PMETB: 'Methods will be chosen on the basis of validity, reliability, feasibility, cost effectiveness, opportunities for feedback, and impact on learning'.[6]

To both of these should perhaps be added the issue of standard setting. This paper considers the characteristics of the specialty certificate examination in terms of each of these in turn.

Validity

Validity asks whether the examination tests that which is intended. It is not possible to quantify this but only to make a judgment on the evidence available. Two types are content and construct validity. Content validity asks whether the examination is representative of the whole domain. An important piece of evidence for this is how well the examination is 'blueprinted' to the curriculum; are all the outcomes tested regularly with a good mix in each paper? The final report on the pilot project aimed to report on validity but did not do so directly.[2] A blueprint was used, however, although it is not clear to which outcome.

Moving on from the pilot, PMETB requires that 'The blueprint detailing assessments in the workplace and national examinations will be referenced to the approved curriculum…'.[6] In response to this there are grids in the curriculum documents (for example, for diabetes and

John Cookson, professor of medical education and undergraduate dean, Hull York Medical School

endocrinology[7]) which map the outcomes of each course to the examination modes (specialty certificate examination, mini-CEX etc). What seems to be missing is a robust method of blueprinting the specialty certificate examinations to the curriculum. How will it be clear that a particular paper samples the content fully? For example, the endocrinology and diabetes regulations indicate the overall percentage that each part of the syllabus (adrenal, thyroid etc) has within the examination, the gastroenterology regulations list 16 disease areas from which the examination will be set.[8,9] However, neither of these lists is congruent with the format of the curriculum outcomes and could not be linked to it in its current form.

Indeed, the current framework adopted for the outcomes would make this problematic. They all have a rather old-fashioned division into knowledge, skills and attitudes while the current view is that this leads to an artificial compartmentalisation rather than an integrated approach. An example concerns peptic ulcer bleeding in the gastroenterology curriculum.[10] This is given as:

- knowledge: 'define the pathophysiology of arterial bleeding endoscopic and radiological diagnosis, endoscopic and surgical treatments'
- skills: 'able to evaluate the indications for urgent endoscopy for diagnosis and treatment of bleeding peptic ulcer'
- attitudes: 'demonstrates willingness to recommend prompt endoscopic action and liaise with surgical colleagues as necessary'.

This might be more usefully expressed as 'Diagnose and manage patients with acute bleeding peptic ulcer with appropriate urgency in collaboration with surgical and radiological colleagues as necessary'.

There is in addition much confusion about what constitutes an attitude. For example, the dermatology curriculum lists 'Recognises dangers of prick testing', which is actually a mixture of knowledge and skill.[11] The attitudes behind it are those of a careful reflective practitioner who weighs the balance of risk and benefit. This is a generic trait (recognised as such in the respiratory medicine curriculum) which cannot be successfully anatomised in this way. These issues are important because if a knowledge-based examination is blueprinted only to the knowledge section then important areas will be missed.

A construct is 'a personal psychological characteristic that cannot be observed directly but which is presumed to exist'.[5] As an example, one might presume that experts would score higher than novices in a test if it was a true representation of their abilities. This is an illustration of construct validity. There are data in the final report that provide some evidence on this because a mix of clinicians took part from specialist registrar (SpR) to consultant level.[2] Happily there was a trend in all four specialties tested for senior doctors to perform better than juniors. Improvements, however, were modest ranging from 5.9 percentage points in geriatrics to 12.4 in neurology which, over a minimum period of six years (five years training + at least one year as a consultant), amounts to between 1 and 2 percentage points per year.

Reliability

This asks about the reproducibility of a test. If the first test, or a similar one, was run again would the results be the same? The usual measure of reliability for written tests of the MCQ format is Cronbach's alpha and the generally accepted figure for a reliable examination is above 0.8 although 0.9 is preferred and is easier to achieve with a written test than with a clinical exam. The final report on the pilot calculated Cronbach's alpha as between 0.55 and 0.81 for the different specialties.[2] However, reliability is highly dependent on the number of questions and the low figure stemmed from cardiology with only 50 questions. The report calculates that the number of questions necessary to reach acceptable reliability would be 200 and this has been accepted for the main examinations.

Educational impact

Candidates work to the examination so the content and format of the examination have a crucial role in determining candidate behaviour. There is, as yet, no information on educational impact apart from some positive comments from candidates on the stimulus to study and some negative comments that the format was inappropriate for clinical practice.[2]

Cost effectiveness

As in medicine, an examination can be educationally effective but not cost effective but if it is cost effective it is also educationally effective. There are very little data on the costs of any medical examinations. The report regrets that it did not formally address the question of costs but they appear to have been substantial and were mainly in terms of consultant time for question writing and standard setting.[2] A key question, therefore, is the number of new questions required each year. These are summative examinations; test security is important but inevitably there will be leaks limiting the ability to use questions repeatedly and knowledge also changes. The fees are currently set at £800.[12]

Acceptability

Any examination system has to be broadly acceptable if it is to work properly. This involves not only the candidates and examiners but those who have to operate the systems and those who rely on the results. It is likely that any process will become more acceptable when it becomes more familiar. The report did collect feedback comments from candidates; about 25% were broadly positive and 75% broadly negative.[2] There were no formal comments from examiners but it is clear that many found it difficult to make the time commitment necessary to write good quality questions.

Standard setting

Standard setting is arguably the most important part of the process. It is little use having an ideal examination framework if the end result is not arrived at by a clear defensible method. Most medical examination systems now adopt a criterion-referenced approach and use one of several methods designed to collate the individual judgments of a number of examiners into a 'cut score'. One of the most common is the Angoff which was used for three of the specialties in the pilot. It is not easy to do a reliable Angoff, one of the difficulties being the identification of the nature of the 'borderline candidate'. This may have been behind the problems in the pilot where the pass mark was variously set at 83% in cardiology (4.8% pass rate) to 57% in neurology (84.5% pass rate).[2]

Discussion

There is as yet limited evidence to allow a full assessment of the specialty specific examination against the criteria given. There are no published data from the examination process proper. The pilot remains a pilot but did, however, have 1,197 participants so its findings are of considerable value and it is possible therefore to discern a number of issues that may need to be more fully addressed as the programme rolls out. It seems that it is possible, based on the evidence of the pilot, to achieve reasonable reliability with an examination of 200 questions (over two papers).

Although reliability is important, a reliable examination is of no value if it does not test what is intended so validity issues need to be considered alongside reliability. This is not easy since there is no test for validity. The evidence so far does, however, raise some questions. To improve content validity there would seem to be a case for blueprinting the examination papers more closely to the learning outcomes rather than just the examination processes to the outcomes. To be successful this may mean moving away from the knowledge/skills/attitudes framework currently used.

The evidence on construct validity is of interest. Why is there such a small difference in marks between entrants to the training programme and consultants? Two possibilities are that the examination is not sufficiently discriminating or that the essential difference between junior and senior doctors (assuming that there is one) is not primarily knowledge acquisition. In fact the findings of the pilot are not altogether surprising considering what is known about the nature of expertise. This may not depend much on knowledge or reasoning skills but on wide experience of relevant patients.[13]

Further, the use of a knowledge-based examination runs counter to the widely accepted principle of Miller that assessment systems which test 'shows how' and 'does' are more appropriate than 'knows' and 'knows how' particularly for professionals at later stages of their training. It would seem important that as the examination is rolled out, more data are collected about candidates at different stages of their training and from consultants but at the moment it cannot be assumed that the examination is a valid test for this group of trainees.

Any examination system has a major educational impact and should therefore be used to promote the learning expected. If more than one type of learning is expected, more than one type of examination system is necessary. Since the overall aim here is to produce competent specialist clinicians, a suite of assessment tools is appropriate. Knowledge assessments will drive candidates to the books and clinical assessments to the bedside so it is important that the relative place of knowledge and clinical experience in a putative specialist is understood and reflected in the examination system otherwise there is a risk of perverse effects on learning.

Determining the cost of an examination is difficult. Unless there is to be a subsidy then there needs to be sufficient fee-paying candidates to support the examination. Numbers of SpRs given in the pilot suggest that there could be around 40 candidates per year in dermatology and 100 in geriatrics (plus re-sitting candidates) and correspondingly fewer in the smaller specialties. Fee income would, therefore, be between £32,000 and £80,000. There is also the opportunity cost to be considered; time and money spent on this examination cannot be spent, for example, on increasing the reliability of a clinical one. It seems probable that taking into account overall costs, the examinations will run at a loss.

No one likes examinations or change so new examinations are unlikely to win many plaudits until they become well established and the metrics calculated and understood. Standard setting in medical examinations has become of major concern; rightly as it will determine success or failure for candidates. The Angoff method or some modification is widely accepted and seems appropriate for this examination but to do it properly and consistently needs considerable time and commitment. It may be relatively straightforward to agree on the 'borderline' criteria for completion of the CCT/consultant appointment, but more difficult after year three of the programme as proposed. It will be important to factor the costs of standard setting into the overall cost-effectiveness model. The examinations will quickly lose credibility if the pass rates for the range of specialties are significantly different. If the modified Angoff as proposed in the examination regulations is to include using the actual results of candidates to inform the process, this may help to prevent discrepancies.

References

1 MRCP(UK). Membership of the Royal Colleges of Physicians of the United Kingdom. www.mrcpuk.org/SCE/Pages/Home.aspx

2 Joint Committee on Higher Medical Training. *Knowledge-based assessment; pilot project*. London: Joint Committee on Higher Medical Training, 2006. www.jrcptb.org.uk/SiteCollectionDocuments/KBA%20Project%20Final%20Report.pdf

3 Joint Royal Colleges of Physicians Training Board. www.jrcptb.org.uk/Specialty/Pages/default.aspx

4 MRCP(UK). Membership of the Royal Colleges of Physicians of the United Kingdom specialty certificate examination regulations and information for candidates. www.mrcpuk.org/SCE/Pages/Regulations.aspx

5 Schuwirth LWT, van der Vleuten CPM. *How to design a useful test; the principles of assessment*. Edinburgh: Association for the Study of Medical Education, 2006.
6 Postgraduate Medical Education and Training Board. *Standards for curricula and assessment*. London: PMETB, 2008.
7 Joint Royal Colleges Physicians Training Board. *Assessment blueprint for endocrinology and diabetes mellitus*. London: JRCPTB, 2007. www.jrcptb.org.uk/Specialty/Documents/Endocrinology%20and%20Diabetes%20Assessment%20Blueprint.pdf
8 Federation of the Royal Colleges of Physicians of the UK. *Specialty certificate examination in endocrinology and diabetes. Regulations and information for candidates*. London: Federation of the Royal Colleges of Physicians of the UK, 2008. www.mrcpuk.org/SiteCollectionDocuments/SCERegulationsEndocrinologyandDiabetes2008.pdf
9 Federation of the Royal Colleges of Physicians of the UK. *Specialty certificate examination in gastroenterology. Regulations and information for candidates*. London: Federation of the Royal Colleges of Physicians of the UK, 2008. www.mrcpuk.org/SiteCollectionDocuments/SCERegulationsGastroenterology2008.pdf
10 Joint Royal Colleges Physicians Training Board. *Specialty training curriculum for gastroenterology*. London: JRCPTB, 2007. www.jrcptb.org.uk/Specialty/Documents/Gastroenterology%20Specialty%20Training%20Curriculum%20May%202007.pdf
11 Joint Royal Colleges Physicians Training Board. *Specialty training curriculum for dermatology*. London: JRCPTB, 2007. www.jrcptb.org.uk/Specialty/Documents/Dermatology%20Specialty%20Training%20Curriculum%20May%202007.pdf
12 MRCP(UK). Membership of the Royal Colleges of Physicians of the United Kingdom. Exam fees. www.mrcpuk.org/SCE/Pages/ExamFees.aspx
13 Norman G. Research in clinical reasoning: past history and current trends. *Med Educ* 2005;39:418.

Address for correspondence: Professor J Cookson, Hull York Medical School, University of York, York YO10 5DD. Email: john.cookson@hyms.ac.uk