Royal College of Radiologists Examinations Review
# Royal College of Radiologists Examinations Review

## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background and Purpose of the Review</td>
<td>4-5</td>
</tr>
<tr>
<td>Review Team</td>
<td>6</td>
</tr>
<tr>
<td>Approach</td>
<td>7-10</td>
</tr>
<tr>
<td>Findings and recommendations (Introduction)</td>
<td>11-14</td>
</tr>
<tr>
<td>RCR central assessment Activity</td>
<td>15-19</td>
</tr>
<tr>
<td>FRCR (Clinical Radiology) Overview</td>
<td>20-23</td>
</tr>
<tr>
<td>First FRCR</td>
<td>24-27</td>
</tr>
<tr>
<td>Final FRCR A</td>
<td>28-30</td>
</tr>
<tr>
<td>Final FRCR B</td>
<td>31-36</td>
</tr>
</tbody>
</table>
# Contents (continued)

<table>
<thead>
<tr>
<th>Section</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>FRCR (Clinical Oncology)</td>
<td>37-39</td>
</tr>
<tr>
<td>First FRCR</td>
<td>40-42</td>
</tr>
<tr>
<td>Final FRCR A</td>
<td>43-45</td>
</tr>
<tr>
<td>Final FRCR B</td>
<td>46-52</td>
</tr>
<tr>
<td>Equality &amp; Diversity/Protected Characteristics</td>
<td>53-55</td>
</tr>
<tr>
<td>Summary</td>
<td>56-62</td>
</tr>
<tr>
<td>Appendices &amp; Bibliography</td>
<td>62</td>
</tr>
<tr>
<td>Team biographies</td>
<td>62-65</td>
</tr>
<tr>
<td>Bibliography and resources</td>
<td>66-67</td>
</tr>
</tbody>
</table>
Background and Purpose of the Review

The Royal College of Radiologists (‘The College’) has a wide reaching scholarly remit to advance education, research and study in the practice of two key clinical disciplines – Clinical Radiology and Clinical Oncology. A major part of this remit is a broad educational portfolio that encompasses teaching, learning and assessment of junior medical staff who are training in both clinical disciplines:

‘The College is responsible for setting the curriculum for its two specialties to ensure that high educational standards are met in the interests of safe and responsible practice, assessment of schemes for training in the specialties of clinical radiology and clinical oncology and defining and monitoring programmes of education and training for clinical radiologists and clinical oncologists at all stages of their careers’

https://www.rcr.ac.uk/content.aspx?PageID=1267

Assessment of the learning outcomes of both curricula is undertaken by a range of written and performance tests within a clearly described programme of assessments (publically available via www.rcr.ac.uk). Demonstration of competency within all stages of this programme leads to the award of Fellowship of the Royal College of Radiologists (‘FRCR’), which is a key requirement for progression of junior medical staff (‘Trainee’). The primary purpose of the FRCR is to assess trainees within UK training schemes, but Fellowship examinations are also undertaken by doctors working in clinical oncology and clinical radiology in international settings, although the award does not necessarily determine career progression in these settings. In support of the international scope of the FRCR examinations, the College delivers both written and clinical performance assessment in the UK and overseas, with some 9 600 members and fellows worldwide.
Within the College, there is an active programme of assessment improvement within both FRCR examinations, and a strong commitment to governance and shared practice (through the activities of the Fellowship Examinations Board). As part of its ongoing improvement ethos, the College sought an external review of the assessment programmes of the FRCR in Clinical Radiology (‘CR’) and Clinical Oncology (‘CO’):

‘To provide an independent, analytical review of the governance, planning, delivery and outcomes of the Fellowship of the Royal College of Radiologists (FRCR) examinations in Clinical Radiology (CR) and Clinical Oncology (CO) to establish their quality assurance mechanisms, fitness for purpose and fairness.’

This review uses the General Medical Council’s ‘Standards for curricula and assessment systems’ and modern principles and practice of educational assessment (e.g. those used by the American Educational Research Association) as benchmarks for the review. A range of core references and support material is contained within the Bibliography.

In doing so, the review team is cognisant that whilst the award of the FRCR is a requisite aspect of a trainee’s progress, it forms part of a wider evidence base that draws on workplace based performance and professionalism assessments. Examination of these latter aspects does not form part of the review.

The broad areas within ‘Standards for curricula and assessment systems’ are:

**Planning**  
Clarity of the sequence and function of assessment system and accessibility of information

**Content**  
Sampling and blueprinting against the curriculum and identified learning outcomes

**Delivery**  
Overall validity, reliability and feasibility of individual assessments and the assessment system

**Outcomes**  
Standards and feedback – for all stakeholders

**Review**  
Resources, lay involvement and equality and diversity

These standards are used to draw together an evidence set that assists the RCR in meeting the remit and responsibility for oversight of its high quality assessment systems.
The Review Team

The review process has been undertaken by a consortium of 4 reviewers, who as a group have considerable national and international expertise in the field of assessment, providing consultancy and academic support to institutions worldwide.

Dr Richard Fuller
Associate Professor & Director, MBChB Programme
Leeds Institute of Medical Education, University of Leeds

Professor Trudie Roberts
Professor of Medical Education & Institute Director
Leeds Institute of Medical Education, University of Leeds

Dr John Patterson
Honorary Senior Lecturer in Medical Education
Institute of Health Sciences’ Education
Bart’s and the London School of Medicine and Dentistry, Queen Mary University of London

Professor Katharine Boursicot
Associate Professor & Head of Assessment
Lee Kong Chian School of Medicine
Singapore

Brief biographies are given in the Appendices and full curricula vitae are available on request.
Approach

Area 1: Assessment practices review

A critical review of the assessment strategy for the award of FRCR (CR) and FRCR (CO) has established the degree to which it is aligned with GMC requirements and the extent to which it follows current best practice. The review also notes (and in discussion with the College, made recommendations) in respect of external access to policy and guidance (for other training bodies, candidates and the wider public) as part of the GMC’s standards on assessment planning.

Area 1 is critical to establishing that the College’s assessments meet modern standards for comprehensive assessment validity. This includes the choice of appropriate assessment instruments to target effectively the attributes (competencies) to be assessed, how the corpus of knowledge, skills and behaviours associated with each competence is sampled and how examiners judge that a satisfactory standard has been achieved by the candidates. The review addresses these questions both at the level of the individual assessment (e.g. FRCR CR Part 1) and across the whole assessment system (e.g. FRCR CO), with focus on programmatic quality improvement.

The review has been undertaken as a mixed desk based and direct practice observation piece of work, and encompassing both pre-assessment delivery activities (blueprinting, item writing and standard setting), conduct and delivery of assessments (particularly oral and clinical assessments) and post hoc processes (outcomes and development). This has involved scrutiny and analysis of over 150 detailed policy documents, analysis of 50 results data sheets, and access to multiple webpages. A number of telephone and face-face discussions and interviews were undertaken with RCR college staff, examiners, senior faculty in both CO and CR examinations and trainees.

A number of components of the CR and CO assessments are undertaken in international centres. The review team did not observe the delivery of assessment in these centres, but examined College policies for standardization, security and equating alongside psychometric analyses of site variation and candidate performance.
The team used a validity based approach to assess the effectiveness of the College’s assessment practices in respect of CR and CO Fellowship examinations. Contemporary validity theory involves a two step process, firstly to generate an interpretative use argument - namely, the purpose of the assessment and the application/inferences that are drawn from the scores obtained. This reflects that validity is a property of the inferences we draw from test scores, and how we use them and that is validity is not a property of the ‘test’. The second phase involves designing a framework to gather evidence (a validity ‘chain’) that supports the use argument. As such, validity can be viewed as iterative, continuous process of quality improvement as each validity review reveals areas for further improvement and development.

In respect of the assessment practices review, this framework will encompass:

- The clarity of purpose of the FRCR (CR) and FRCR (CO) assessments
- The choice of assessment formats and supporting evidence
- The content, item selection, blueprinting and scoring formats
- Response formats (including assessor training, behaviour and feedback)
- Whole- and item-level test metrics (the issue of protected characteristic monitoring is detailed in Area 2 of this proposal)
- Predictive validity – and correlation with other assessment/performance data held about candidates
- Consequential validity, namely the FRCR’s policies on the consequence of success on trainee progression, and the impact of failure (including feedback and further attempts at assessment)

A major component of the review focuses on measurable aspects of quality such as test reliability (the internal consistency and reproducibility of candidates’ scores) and psychometric quality of single test items (e.g. SBAs in knowledge tests, or ‘stations’ within performance assessments). These whole-test and item-level analyses not only establish the accuracy of, and confidence in test outcomes, but also form an important quality loop to improve the performance of examinations in the future.
As a product of this element of the review, we have been able to highlight a significant number of areas of excellent practice across the assessment functions provided by the College, CR and CO examinations. Based on the evidence gathered from this validity framework, the report makes a number of recommendations in respect of the basic structure of the assessments contributing to the FRCR awards, pre- and post hoc assessment practices in the college and the augmentation of existing quality assurance processes. This is covered in depth in the Findings and Recommendations section.

Area 2 Assessment outcomes review – focus on protected characteristics

Within the ‘Review’ section of Standards for curricula and assessment systems are requirements that equality and diversity monitoring is an established part of any comprehensive system. A concern for many postgraduate medical qualification awarding bodies is the extent to which UK medical graduates (and/or those training in the UK) may perform differently to international graduates/those training overseas. This concern has generated ongoing research, focusing on gender, age, ethnicity and country of qualification.

A full-scale study would be both complex to undertake and difficult with the relatively small numbers of FRCR candidates, so for the purposes of this review, the team has focused on investigating the quality of the College’s data systems to capture performance by protected characteristic, and make simple comparisons between UK and IMG candidates’ performance in the various parts of the award. However, the team recognises that the multiple confounders that contribute to candidate performance are complex, nuanced and care must be taken in both analysis and interpretation – particularly given that elements of the FRCR awards are examined in international centres. The college cannot be responsible for the international candidate population that presents (and selects) itself for examination at a particular time, but must ensure as far as possible that all candidates are treated fairly within an assessment system that is designed to identify and pass only those candidates who are competent.
The review team will also examine the composition and policy surrounding key appointments by the RCR to assessment boards and committees. In doing so, the review team will review arrangements to support professional development of current and future members of these boards, as a marker of the RCR’s policies to ensure expertise and appropriate representation of the diversity of its membership.

Whilst the primary purpose of the review was not to carry out systematic literature searches on topics of interest the College, a bibliography relevant to criticisms and recommendations contained within this report is available at the report.
Findings and Recommendations

The report will provide a summary of both FRCR assessment systems, a focus on the assessment support provided by the College, and then a review of each component examination under the following themes:

a) Areas working well (including good practice that can be shared within each specialty as well as across specialties via the Fellowship Examination Board).

b) Psychometric quality report (a summary of key findings, with full reports provided separately).

c) Areas for development and recommendations (including resources and faculty development opportunities).

With each recommendation, a tag will relate to the GMC Standards framework as follows:

- **P** Planning of the assessment system/availability of information
- **C** Content and blueprinting
- **D** Delivery and whole system/component part quality metrics
- **O** Outcomes, standards and feedback
- **R** Review, resources, governance, equality and diversity
Summary of assessment systems

FRCR (CR)

The clinical radiology system comprises of eleven components, spread over three parts. The assessments are conducted in parallel in the UK (based across a number of sites including the College) and in two overseas venues.

First FRCR (CR1) Two part knowledge test using recognition of anatomical clinical images with a ‘name the structure approach’ and a true/false format multiple choice question (MCQ) to assess physics knowledge

Final FRCR A (CR2A) Modular assessment of 6 curriculum modules using Single Best Answer (SBA) format questions

Final FRCR B (CR2B) A tripartite assessment comprising ‘Reporting’ and ‘Rapid Reporting’ formats, and a case-based oral examination.

There are clear rules in respect of entry to the First FRCR, and candidates must be in a recognized training post. For candidates within UK training schemes, progress through FRCR is carefully aligned with expected in-training progression, with clear statements in respect of the consequence of failure within the FRCR (extensions of training time and withdrawal from training schemes).

Candidates must progress through the sequence of assessments, although they are permitted to take both the CR1 elements and the modular components of the Final FRCR A in any sequence, however these must be (a) completed within a limited number of attempts and (b) all be passed before progressing to the Final FRCR B. Admission to the Final FRCR B is also dependant on satisfactory completion of an approved period of Specialty Training, determined in the UK through workplace assessment, supervisor reports and a satisfactory Annual Review of Competence Progression.
Candidates must pass each part of the assessment independently without compensation. There are clear non-conjunctive scoring rules in place within the Final FRCR B to ensure candidates who pass achieve satisfactory performance in all elements. The CR assessment system has discontinued an unlimited number of candidate attempts and now has a clearly described limited attempts policy in place, recognizing GMC guidance in this area.

A considerable amount of structural guidance (including for candidates and examiners) is freely available on the College website, and reflects commendable practice. This includes clear information in respect of the College’s appeals processes.
FRCR (CO)

The clinical oncology assessment system comprises of seven components, spread over three parts. The assessments are conducted either in the UK (in the College for the First and FRCR A and clinical sites and the College for FRCR B) or at another international centre (currently in Hong Kong).

First FRCR (CO1) Modular SBA examination (4 curriculum modules)
Final FRCR A (CO2A) Integrated SBA examination
Final FRCR B (CO2B) Clinical Performance and structural oral examination

There are clear rules in respect of admission to the First FRCR and progression through the assessment series (including rules in respect of completed clinical oncology training for entry to the Final FRCR). For candidates within UK training schemes, progress through FRCR is carefully aligned with expected in-training progression (as determined by Annual Reviews of Competence Progression, with clear statements in respect of the consequence of failure within the FRCR (extensions of training time and withdrawal from training schemes).

Candidates must pass each part of the assessment independently without compensation. There are clear non-conjunctive scoring rules in place within the Final FRCR B to ensure candidates who pass achieve satisfactory performance in all elements (clinical and oral). The CO assessment system has discontinued an unlimited number of candidate attempts and now has a clearly described limited attempts policy in place, recognizing GMC guidance in this area. There is a clear system of learner appeals.

This information is clearly laid out as part of the overall assessment strategy for clinical oncology training alongside support for candidates and examiners, and is publically accessible via the College website, which is commendable practice.
Assessment functions undertaken by the College

The College undertakes a range of central functions to support the delivery of the CR and CO Fellowship examinations. This is undertaken by a dedicated team of an examinations manager, project officer and administrative staff who support both the delivery of assessment and image/question banks. A range of other functions are provided by the team to support high quality assessment operations – examiner recruitment, candidate registration, assembling examination papers, undertaking automated scanning/marking, collation of performance data and quality metrics. Some of the team are critical members of a new ‘Automation’ project which will further enhance the banking, tagging and selection of high quality images in the FRCR CR examinations in particular (http://www.rcr.ac.uk/docs/about/pdf/RCR(14)6_Annual_review_20132014.pdf).

The RCR assessment review necessitated significant working with the exams team, both through our desk based work and observation of assessments delivered at the College and on clinical sites. We observed assessments to be delivered in a highly efficient and professional manner with excellent arrangements to support examiners and candidates. The exams team support examination boards, which were also conducted in a professional and rigorous manner.

A number of areas of College function are notable for excellent practice:

1. **Organization and security.** Clear arrangements are in place to maintain the security and currency of assessments despite the complexity of parallel international centres. Time delays, candidate ‘quarantine’ arrangements and different papers are deployed proportionally to avoid leakage of confidential assessment material.

2. **Invigilation and Incident Management.** There are clear policies in respect of these functions, including meticulous attention to detail in the CO Clinical examinations.
3. **Externally facing guidance to candidates.** Practice papers made available online including YouTube instructional videos and sample response forms. Particularly noteworthy is the approach by the CO examinations whereby older questions are continually retired and made available for candidates, although care should be taken to ensure they remain of an acceptable quality.

4. **Board Governance structures.** There is a well described structure in respect of the roles and population of examination boards, including the recruitment of examiners for all parts of the assessment process. Whilst the review did not look specifically at equality and diversity issues within the examiner pool, there appears to be diversity of examiners across age range, gender, ethnicity and geographical distribution within the UK evident in some boards.

   Similarly noteworthy is the open access publication of reports from components of the CO examinations about the general conduct of assessment, candidate performance and relative strengths and weakness. This is unusual and clear evidence of the College’s approach to aspects of the ‘Planning’ theme within *Standards for curricula and assessment systems*.

5. **Reasonable adjustment policies.** These are clearly described, with an extra time allocation/additional adjustments based on recommendations from disability advisers and/or supporting evidence from Training Programme Directors. This applies across all of the FRCR examinations.

6. **Standard Operating Procedures.** These are clear and detailed, and cover a range of essential functions of the exams team (production of question papers, analysis and results handling, scanning and cross-checking procedures).

7. **Appeals.** There is publically available guidance in respect of how the College manages appeals from candidates. During the review (and discussion with examiners) there is also clear evidence that the College is a learning organization, based on improvements to processes after previous learner appeals.
There are a number of overarching areas for development across the whole assessment system, with consequent recommendations:

1) **Strengthening the remit and responsibility of the Fellowship Examination Boards (FEB).** The College has made good progress in the creation of overarching boards to manage what might be considered a ‘federal’ structure of examination boards and governance structures for individual test components. A key feature of the review, and discussion with examiners was the relative lack of integration between certain examination components, manifested through poor understanding of purpose and how each component contributed to an overall assessment system. Examiners often viewed the Final Part B clinical examinations as a ‘true’ test of acumen, and were unable to describe the value of the knowledge tests.

Trainees spoke highly of the RCR examinations, the helpful nature of the College, examiners and the FEB as a central point for contact – but raised concerns in respect of variable pass rates between examination components, standards and ‘fairness’. The FEB has a central role in ensuring oversight and a ‘helicopter’ view of overall assessment systems, candidate progression and consistent standards to help preserve the currency of the Fellowships.

We would also encourage the College to continue facilitating the Joint FEB meeting as a way of sharing good practice and development of assessment. Consideration should be given to the membership of the FEB to include other key stakeholders (e.g. employers and lay members).

2) **Developing a central repository for relevant CR and CO documentation.** Whilst the majority of documents required for this review were easily available, it was clear that critical documents were often held at the level of individual examination boards without knowledge of the College. In an environment of increasing learner appeals and challenges to assessment, such a system presents a risk to both FEBs and the College if key documentary evidence is not readily available.
The review team suggests that as by-product of this review (and the gathering of a full set of evidence), that the College and FEBs develop systems for documentary governance, with core documents and evidence held centrally by the College secretariat, with a system of regular review and responsibility for this documentation.

3) **Improving mechanisms for data capture.** The College and examination boards hold sensible datasets in respect of candidate scores and achievement in individual examination components. This has allowed the review team to undertake an in-depth psychometric analysis which is detailed in the following commentary and separate reports. Longitudinal tracking of cohorts and item performance in some assessment components is already established practice.

However, there is limited data available to perform additional analyses, which would the help the College enhance its evidence set for the Outcomes and Review sections of the GMC’s standards in the following areas:

- Longitudinal tracking of candidate performance and progression through the CR and CO assessment diets.
- Routine measures of non-candidate variance factors (examiner pairs, sites, times when candidates sit assessments).
- Ensuring routine monitoring of candidate protected characteristics (although the Review Team note that this is contingent on candidate self-declaration).
- Recording and monitoring of assessor characteristics (including gender, appointment, equality & diversity training and geographical distribution) to allow demonstration that the CR and CO examiners are fully representative of the corpus of senior clinicians who are eligible to examine.
As a minimum this should generate a database that contains the performance (including final marks) for each candidate's assessment history while taking College examinations. This should also include:

- The resit history and resit scores for every candidate.
- A record of candidates' scores on every item and their overall score for every examination.
- Clear linkage between the candidate number for each examination and their unique identifier in the collective database.

Such a database will allow longitudinal studies of candidates across their various examinations and will in time generate sufficiently large pools of candidates to compare the performance of candidates by protected characteristic, resit attempts, gender and place of primary medical qualification and training. It will also assist the College in triangulating any concerns from trainees in respect of their training and passing rates by Deanery.

At present, basic examination statistics are provided for each single best answer MCQ examination when the papers are optically read by Speedwell software. A simple Excel-based system (as used in this report) could be introduced which could then be applied to all assessments, including clinical examinations. This would provide easy to digest information that can be fed back to examiners and question writers as part of quality improvement.
FRCR Clinical Radiology

Overview

The FRCR (CR) is a well established assessment system that uses a variety of formats designed to test clinical knowledge and performance/skill. Whilst each component part will be discussed in detail in the following pages of the report, the review team felt a number of broader areas of practice merit commendation, and provide springboards for further improvement within the CR assessment system.

The curriculum for CR is clear with a number of very well laid down learning outcomes, and clarity in respect of how elements of the CR1 and CR2A examinations map directly. This is further reflected in the significant improvements made in terms of tagging questions and developing the bank of questions (particularly in respect of the CR2A). Support and guidance for question writers and examiners across the Final FRCR was particularly noteworthy – and the review team commends the innovative, authentic approaches to assessment which include the Reporting/Rapid Reporting elements of the CR2B.

There are a number of overarching areas for development across the whole assessment system, with consequent recommendations.

1) Development of a ‘use argument’. Whilst statements such as *this component of the exam is designed to test* ….’ do exist within documentary evidence, they are often hidden away in examiner or question writing guidance, rather than in externally facing communications to candidates. Clarity about the purpose of test, the level of training of candidate and the application of results is an important mechanism that justifies the choice of assessment, test format and spread of the assessment sample. Development of such statements will assist in agreeing whether e.g. CR Physics should be test of knowledge or knowledge application, and e.g. how the CR2B focuses on important aspects of practice that cannot be reliably assessed by workplace assessment methods.

**Recommendation**: That the CR examination produces a clear statement of purpose for the Anatomy and Physics components and makes this publically available to all stakeholders.  

(P)
2) **Blueprinting.** The review team found that the term ‘blueprinting’ was interpreted in a variety of ways in the CR assessment process, most often to indicate that a particular question had been tagged against the CR curriculum. Tagging is an area of strength for much of the CR, and basic decision rules exist to guide equal spread of papers by curriculum content, but no definable examination blueprints exist.

Blueprinting is a process by which the assessment is mapped directly against the curriculum learning outcomes, allowing appropriately broad sampling, a focus on areas of importance as well as decision rules that dictate what proportion of the test relates to different areas of the curriculum. This is typically expressed as a grid format (or some form of matrix with axes) and is essential in the justification and defence of test content. In assessment systems with multiple components, blueprints can be ‘stacked’ to ensure a candidate’s assessment journey can be tracked. This is particularly important when individual exam boards are defining assessment content in isolation, and will assist the FEB’s overarching view.

**Recommendation:** That the CR examination generates blueprints routinely as part of assessment planning. Examples of helpful clinical test blueprints exist within the Final FRCR A in Clinical Oncology or on the GMC’s website. (C)

3) **Test Standardisation and equating.** Akin to the majority of major educational organizations undertaking multi-site assessment, the CR examination uses multiple different diets of papers to help with security. However, there is no mechanism to identify whether a paper is of a similar standard to others, other than broader performance data. Given the lack of blueprint, it does risk producing a paper that may well differ significantly in difficulty, but is not compensated for by approaches to standard setting.

Score linking is a process whereby different papers (e.g. for different diets) are constructed to the same content and statistical specifications. Equating involves using the raw scores from these different papers onto a common scale, and using small statistical adjustments to account for differences in the difficulty of each paper. This allows the deployment of multiple papers over time, with the ability to yield scores that can be sensibly interpreted. Deploying multiple different papers also increases test security.
**Recommendation**: A simple approach would be to nest a series of anchor questions across papers to track whether performance data is consistent across different diets of candidates, or if not, seeks to understand this (e.g. if there is evidence that the different sittings of the exam are undertaken by candidates of different seniority). More sophisticated approaches include longer term moves to using test equating (e.g. by application of Item Response (IRT) theory based mechanisms), a technique used by other large educational institutions.

4) **Standard setting approaches.** Established methods of standard setting exist for the CR1 and CR2A, using the modified Angoff method and application of the Hoftsee method. Formal, criterion referenced standard setting is an essential aspect of validity evidence, and the CR exam use clearly established methods with evidence of examiner training. Application of the Hoftsee method is an area of significant debate, as it was not originally conceived for high stakes testing. However, an argument can be made that application of the Hoftsee (a compromise method to manage passing rates) can be of value in ensuring ‘sensible’ passing rates, given how tightly linked the FRCR CR is to training progression and workforce planning. Of note the Hoftsee method is used by other postgraduate medical institutions for similar reasons.

Of more concern is the fixed passing standard that is applied to the CR2B. This is in part related to a relatively ‘closed’ marking system that is discussed in the Part B component section later in this report.

**Recommendation**: The review team supports the ongoing efforts by the Fellowship Examination Board to review standard setting approaches, including the provision of resources to assist improved engagement with Angoff. Improvements to the process should include a clear articulation of the characteristics of the ‘just passing’ candidate – and a similar approach used in CO2A might be adopted (where there is a clear description of the just passing candidate, which is continually reviewed at each standard setting meeting). As part of the wider review of the CR2B examination, attention should focus on using recognized methods of criterion based standard setting, which is likely to require a review of current scoring processes.
5) Feedback. There is a clearly established literature in respect of the impact (and effect size) on learner improvement and development when provided with high quality feedback. As such, feedback is a key part of the validity evidence chain, forming part of the response process to candidates.

The CR2B produces more detailed feedback to candidates after two attempts at the examination, with a focus on areas of strength and weakness to candidates and their training supervisors. Unfortunately, for the CR1 and CR2A examinations, it appears that candidates receive no feedback beyond whether they have passed or failed. This limits the validity of the assessment process, both for failing candidates and those who are successful – where provision of more individualized feedback will help their own learning, workplace activity and preparation for further assessments.

Recommendation: For the CR1 and CR2A examinations, the current system of automated marking and tagging against the curriculum should allow more thoughtful feedback provision to all candidates, based around candidates’ strengths and weaknesses in different areas of the curriculum. The good practice described for the CR2B examination should continue, but thought needs to be given to the provision of candidate feedback after the first fail in this assessment.

6) Quality loop. A key part of any assessment system is the question ‘how well are my assessments working?’ This question is partly answered through this review process as a significant weight of evidence has been gathered, but modern assessment systems typically rely on a mixture of psychometric and narrative analysis of whole-test and item-level quality data. This often takes the form of overall test reliability (internal consistency), correlation and item level metrics (e.g. discriminant function, test reliability with item deleted). This is typically reported to examination boards, with action plans that detail planned improvements in item design, support for examiners and candidates.

Recommendation: Routine generation of psychometric quality reports for the CR examinations should be produced and reported, as per the suggestions in Dr Patterson’s overview. An extensive range of resources in support are listed in the bibliography.
FRCR Clinical Radiology Part 1 (CR1)

This is a two part examination:

Anatomy: 100 clinical radiology image items (using a ‘name the arrowed structure’ approach). This is hand marked.

Physics: 40 True/False questions with 5 items per ‘stem’ giving a total of 200 items. This is scored an optically read mark sheet with automated marking.

There are established CR1 boards for Anatomy and Physics which is responsible for question writing, setting of papers, examiner recruitment and training, setting standards and routine examination board functions. They are governed by the recently established overarching FEB.

The following comments should be read in conjunction with the good practice and areas for development/recommendation in the CR overview.

Areas working well

1) Test delivery and security. The Physics examination is delivered across a number of diets and venues, and there are clear and comprehensive processes in place to enable smooth delivery and maintain test security, as noted in the areas of central College function.

2) There is a clear purview of the curriculum in relation to anatomy and physics, particularly embedded in question writing and examiner guidance.

3) Clear governance arrangements in relation to examiner selection, support for item writing and the structure and function of the CR1 examination boards. The boards review a number of post-hoc metrics, including measures of internal consistency/reliability (as expressed by the Kuder-Richardson coefficient) to help judge the quality of the examination.
Psychometric Commentary

Full analyses and discussion of the Part 1 CR examinations held in Spring 2014 are provided in the separate ‘Statistical Report on Fellowship Examinations for Clinical Radiology and Clinical Oncology’ and associated annexes. In summary both Anatomy and Physics components perform well in psychometric terms, although there are some potential weaknesses.

The Anatomy paper is run as four separate ‘sets’, raising questions about comparability of paper difficulty and of passing standards. The Physics paper uses true/false MCQ items, which have fallen out of use in recent years because of difficulties in constructing high quality forms of this type of question which tend to test knowledge recall. They have been superseded by single best answer (SBA) formats; particularly as such questions are better at testing knowledge application.

The examinations in Anatomy in Part 1 are given as four different ‘sets’ or papers, which are of the ‘name the structure’ format and are hand-marked. These sets differ in content and probably also in item difficulty. While difficult to use formal test equating procedures because of the relatively small numbers of candidates, the pass mark for each set should be established by a defensible standard setting procedure to ensure that pass-fail decisions reflect as far as possible the actual ability of the candidates assessed. Reliability of these papers is generally good.

The Part 1 Physics examination also performs well with high reliability, although using an outmoded test format as discussed above. Other Part 1 papers have not been examined to see if they test reasoning skills and the application of knowledge. However, testing of cognitive skills at the application level would be appropriate, even at Part 1, in postgraduate assessments.
Areas for development/recommendations

1) **Test format and purpose.** A key component of the development of overarching ‘use statements’ for the CR involves a justification of what level of attainment is being assessed. The review team noted both a feeling from CR2 examiners and in CR1 documentation that CR1 reflected ‘factual recall’ rather than application of knowledge as might be expected for the candidates’ stage of training.

The use of True/False questions within the physics paper represents an older form of knowledge test format, with a well published literature in respect of concerns regarding gender related variance and guessing. Additionally, it is difficult to generate clinically relevant questions where answers are always true or false. The hand marking of the CR1 anatomy paper poses risks in terms of quality and management of datasets as hand marked responses are transcribed and calculated. The College is making significant investment in a new ‘Automation’ project to manage image banks. This enhancement will include generating an online response/computer testing system to overcome the inherent risks in the current process. This will also help the College consider a move to a knowledge application test format for the Anatomy images assessment.

**Recommendation:** That the CR1 Physics assessment should consider a move to a more defensible and recognized test format using Single Best Answer (SBA) formats as in the CO1 Physics assessments and elsewhere. The review team recognizes that writing high quality assessment items for complex scientific material is challenging, but a range of advanced writing workshops exist to provide this support, or could be commissioned by the College as part of faculty development, although it is notable that the CO examination produces SBA format physics questions. An overarching use statement should be generated that is clearly available to all stakeholders.

(P, C)
2) **Blueprinting.** Both assessments produce lists of question topics that are linked to sections of the curriculum, but there is no evidence of formal blueprinting or alignment to core learning outcomes. Coverage and spread of each assessment is reliant solely on examiners’ knowledge setting the papers at the time, with no historical record to guide comparisons of paper difficulty over time, or use of anchor questions to look at comparability of candidate performance and standards.

**Recommendation:** That formal blueprinting is undertaken for each future CR1 examination. As noted in the CR overview, different approaches to blueprinting have been tried by the CO2A examination, and a range of published and workshop based activity to help develop blueprinting are widely available.

(P, C)

3) **Appropriate use of post-hoc analysis data.** Currently, the CR1 boards receive item level statistics derived from a standard statistical software package, comprising point-by-serial, discrimination and quintile performance functions. Unfortunately, the True/False Physics format cannot be analyzed in this way as all 5 ‘items’ in the stem must be answered as true/false – there are no distracting options as would be seen in a SBA format (where this statistical package is correctly used to generate meaningful post hoc data).

(D)
Final FRCR Clinical Radiology Part A (CR2A)

This is a six component ‘modular’ assessment, entry of which is contingent on success in the CR1 assessment. Each paper is directly linked to a major component of the CR curriculum, and takes the form of a 75 item SBA paper. Although candidates can choose to take a variable number of modules at any sitting, there are sensible progression rules in place, including a limit on the number of attempts.

There is an established CR2A board which is responsible for question writing, setting of papers, examiner recruitment and training, setting standards and routine examination board functions. It is governed by the recently established overarching FEB.

The following comments should be read in conjunction with the good practice and areas for development/recommendation in the CR overview.

Areas working well

1) The assessment uses an established test format (SBA) with a continual drive to generate newer, high quality questions. Any one examination uses a maximum of 50% established test items, and there is evidence of policy decisions to target question writing in areas of established deficiency in the question bank.

2) Item tagging against the curriculum is well established – and should be seen as an important step in the springboard to test specific blueprints as noted in the CR overview commentary.

3) Excellent structural guidance to support the genesis of new questions, with clear emendation processes.

4) Support and guidance for examiners. There are clear recruitment policies in place, and a good system of ‘on the job’ training with shadowing/observation mandatory for new examiners. There is a compulsory face-face workshop for new examiners, with plans for enhancement of formal training underway. Question writing guidance for the production of SBAs is also excellent.
**Psychometric Commentary**

Full analyses and discussion of the Part 2A CR examinations held in Spring 2014 are provided in the separate ‘Statistical Report on Fellowship Examinations for Clinical Radiology and Clinical Oncology’ and associated annexes.

The Part 2A examinations in Clinical Radiology take the form of six, separate 75-item SBA papers covering the six modules of the Part 2 award. Reliability, as judged from Cronbach’s α is around 0.72 to 0.78. While just about acceptable, these values are low for well-designed SBA assessments, and higher reliability is desirable for a high stakes assessment such as that for the final award of the FRCR. The notional failure rates for individual papers as calculated from the raw pass marks supplied to the reviewers are very high (35.8% to 65.6 % of the cohort) and the effects will be additive over the six papers.

Candidates can sit the module-based assessments over a period of time and relatively few candidates take all six at the same session. While spreading the time available to complete the six assessments may allow educational development to occur and to be monitored, it may be advisable to have a more synoptic assessment as the award of Part 2 is reached to ensure that knowledge and related skills have been integrated and have not decayed since the individual assessments were first taken.

One Part 2A paper (Cardiothoracic and Vascular) in October 2014 has been seen as part of this review. This paper uses the single best answer format and include clinical scenarios that require demonstration of clinical reasoning. These are generally of a good standard, although some use a ‘which is the least likely’ construct, which is akin to true/false questioning. Such items fail the ‘cover the options test’ which is the first step in the quality assurance of single best items questions. This construction should be avoided.
Areas for development/recommendations

1) **Test format and purpose.** As noted in the CR1 commentary, an overarching ‘use argument’ is a necessary component of modern assessment systems. To a large extent, elements of this use argument – in respect of the purpose of CR2A examination testing knowledge application – are present within question writing and examiner guidance. The review team appreciates that the modular format of the CR2A replaced a more integrated type assessment, but many examiners were not clear about the rationale of the change. As part of this wider review, there is a key opportunity to re-visit the benefits of whether a modularized CR1 with a more synoptic/integrated CR2A better fits the purposes of the examination

**Recommendation:** An overarching use statement should be generated that is clearly available to all stakeholders.

(P, C)

2) **Blueprinting:** Whilst decision rules exist in respect of new: old question distribution, the review team was provided with no evidence of formal blueprinting or alignment to core learning outcomes for each of the modules. Consequently, coverage and spread of each assessment is reliant solely on examiners’ knowledge setting the papers at the time, with no historical record to guide comparisons of paper difficulty over time, or use of anchor questions to look at comparability of candidate performance and standards.

**Recommendation:** That formal blueprinting is undertaken for each future CR2A examination. Such an approach would also allow the CR assessment system to produce a ‘master’ blueprint that gives a helicopter view across all CR2A papers to give better integration. As noted in the CR overview, different approaches to blueprinting have been tried by the CO2A examination, and a range of published and workshop based activity to help develop blueprinting are widely available

(P, C)
Final FRCR Clinical Radiology Part B (CR2B)

This is a three component assessment, with some unique constructs that are well aligned to workplace skills and activity. There are two reporting elements, namely 'standard' reporting items (six cases in 60 minutes, with more detailed traditional reporting) and 'rapid' reporting (35 cases in 30 minutes, candidates are required to note whether abnormal or normal and as appropriate a diagnosis). All clinical images/items are held centrally by the College and subject to review, emendation or rejection after each round of assessment.

Both reporting elements take place at the College. The reporting elements were observed by Dr Patterson who commended the organization, invigilation and delivery of the CR2B. Candidates received an excellent briefing, and had time to practice items in front of their PC/monitors before the examination started – and spare PCs are available in case of equipment failure. Each paper is hand marked and scaled down against a somewhat 'closed' 4-8 algorithm. There appears to be no clarity of purpose behind this scaling.

There are clear non-conjunctive rules that require candidates to achieve a minimum standard in both reporting elements and orals, in addition to securing an overall passing score.

The oral examinations are traditional in format, with pairs of examiners taking candidates through a number of clinical cases. The material for the clinical cases is 'owned' by examiners who bring their own images to the College for the oral examinations, albeit required to conform to a range of cases which fits the modular structure of the curriculum. Each candidate is examined by a pair of examiners for 30 minutes, before rotating on to a second examiner pair. The examination is delivered in a series of sound proof rooms, with candidates rotating between the two pairs of examiners.

Examination facilities are excellent and the oral assessment was attended by Drs Fuller and Patterson. The examination proceeded smoothly with good academic and administrative support. Examiners were enthusiastic and engaged, although (as anticipated in a relatively unstructured oral), we saw significant verbal and non-verbal behaviours directed towards candidates which appear to prompt and miscue.
After each candidate has departed, there is a brief ‘wash up’ period where examiners discuss the candidate’s performance and confirm individual marks, although there is no requirement for agreement within the pairing. Again, marking is performed on a ‘closed’ scale within the 4-8 range, with ½ marks used to generate a numerical mark.

The total marks are combined (reporting, rapid reporting and two examiner pairs). This means the oral examination is weighted, and a maximum of 32 marks are available. There is a pre-determined pass mark of 75% (24/32) with clear non-conjunctive scoring rules to ensure candidate achievement is spread across all components.

There is a dedicated examination board for the CR2B, which meets regularly and is also responsible for examiner training and recruitment, as well as executing traditional functions after assessment diets have completed.

The following comments should be read in conjunction with the good practice and areas for development/recommendation in the CR overview.

Areas working well

1) Authenticity of assessment is a key construct behind elements of the CR2B. This was particularly noticeable in the reporting and rapid reporting components, with the latter attempting to simulate ‘hot reporting’ using the range of common plain images (chest and trauma) and a number of normal films.
2) There is excellent examiner training and guidance, including a comprehensive induction for new examiners which involves observation and shadowing more senior examiners. There is also a nicely described ‘pathway’ with different tasks and responsibilities for examiners occurring at different points of seniority.
3) The CR2B was delivered smoothly and efficiently with good briefing and support for candidates and examiners. Whilst the IT risks associated with examiner ‘owned’ material could be considerable within the oral examination, considerable preparation and testing ensured candidates experienced a smooth exam.
4) Feedback to candidates after the CR2B is thoughtful and detailed for those candidates who have failed twice (and is made available to training supervisors). This could be enhanced further by being made available after the first fail.
Psychometric Commentary

Full analyses and discussion of the Part 2B CR examinations held in Spring 2014 are provided in the separate ‘Statistical Report on Fellowship Examinations for Clinical Radiology and Clinical Oncology’ and associated annexes.

Part 2B comprises three separate elements. There are 30 ‘rapid report’ tasks plus 6 ‘report’ tasks. In both cases candidates take one of four possible ‘sets’ of assessment. While this situation is probably unavoidable, the points mentioned in concerning Part 1 Anatomy should be borne in mind. The third element is two oral examinations where a (sometimes variable) number of cases are discussed. The reliability (Cronbach’s α) of 0.57 for Part 2B overall and for each element (Range: 0.27 to 0.65) is lower than is acceptable for high stakes assessments. Admittedly, it can be difficult to obtain high reliability in clinical assessments, but efforts should be made to improve the reliability of these important assessments. Solutions could lie in two areas – marking schemes and sampling.

A 30-item test such as the rapid reports should be capable of high reliability from the sampling standpoint, so improved reliability should be sought in the way the individual items are scored. Alternatively a greater number of items in the test, if feasible, might improve reliability.

For both reports and orals, the marking scheme could be widened and final marks not compressed as in the current scale – recommendations for improvement are discussed below.

Giving separate scores for each case in the orals, and standardising the number of cases more effectively would be beneficial to the scoring process and using each case’s scores (rather than the overall oral scores) in reliability calculations might give much better reliability indices. The format for the orals used in the Clinical Oncology examinations could usefully be considered for CR2B.
Areas for development/recommendations

1) **Purpose of assessment and test format.** As noted in the main CR review, generating an intended use statement is a critical part of assessment validity. This is particularly important in the justification of test format within the CR2B, where arguments should be made that strongly support the reporting elements of the series, and justify the use of the oral examination. There is clear information within the 2B examiner handbook, and a strong emphasis from discussions with examiners that this exam had a strong focus on clinical safety.

The review team accepts the value of clinical oral examinations for assessing key aspects of performance in highly specialized training fields such as Radiology. However, the oral examinations in their current state cannot be recommended as good practice, because their format and construct allow significant examiner related variance, both in terms of case selection/examiner ‘owned’ material, and examiner behaviours within stations. This poses a number of significant risks to the robustness and defensibility of the oral examination.

**Recommendation:** the CR2B oral assessments need to move to a more structured format that seeks to mitigate these risks, whilst preserving the important, positive features of the oral examination. Adopting a model similar to the CO2B would allow sharing of good practice within the College. A possibility for the CR might be to start each case with mocked up clinical request card – helping preserve authenticity and ensuring the candidate is not left to second guess what the examiner is asking.

2) **Blueprinting and case selection.** Whilst there are some simple decision rules for selection of cases for the reporting and oral examinations, there is no specific test blueprint. The current system is rather reliant on the skill and behaviour of examiners, and lacks the ability to provide an overarching view of the sampling across all three components. This puts the CR2B assessment system at risk – both in case of appeals as well as a lack of integration with those elements tested in CR1 and CR2A. Whilst the images used in reporting are held centrally by the College, there is a reliance on examiners to internally police their own material in the orals (and no way of defending challenges that might lead from this). There appear to be no anchor cases used across examiners pairs.
**Recommendation:** As noted within the wider CR review, there is an urgent need to develop and record test specific blueprints. Doing so should be accompanied by a move that sees all oral cases centrally written and images held by the College. This not only improves the defensibility of the examination, but reduces unnecessary variance *between* stations and helps reduce some of the examiner-related variance factors *within* stations.

3) **A lack of standard setting** is a major concern for the CR2B examinations. Accompanying the comments about scoring and response systems (see below), the lack of a criterion referenced standard for CR2B poses further risks to the defensibility of the examination. It is difficult to support a fixed passing score of 75% with so much potential variance across test material, examiner pairings and candidate factors. Examiners raised additional questions about the appropriateness of some of the non-conjunctive rules within the CR2B (e.g. having to score 75% in each of the components as well as obtaining the overall passing score).

**Recommendation:** The CR2B needs to move to a system of recognised standard setting. The reporting and rapid reporting elements could easily be scored with the Angoff method, and sensible non-conjunctive rules built within this. The orals, if moved to a more structured format as noted in the CO2B could also use a Borderline method for standard setting, which would make use of the supporting statements in the 2B examiner guidance in respect of adequate candidate performance. Borderline methods have the greatest evidence for successful use in performance assessments, but typically become less stable if the total candidate number/assessment diet drops below 50. Angoff methods are also recognized as suitable for setting standards for small cohort performance tests. A range of support resources are detailed in the bibliography.
4) **Scoring mechanisms.** Currently, significant amounts of unintentional and intentional weighting take place using the ‘closed’ 4-8 assessment scale. Constraining examiners to such a tight scale can make it challenging to *numerically* distinguish between candidates, even though examiners were clearly able to recognise failing, borderline and excellent candidates in the examination. We note that the use of ½ marks is employed, but a move to a series of global anchors employed as letters would be more appropriate to avoid some of the well-researched impact of using numerical scores on scoring sheets.

Such a narrow scale is further complicated with the overall totals and scaling (giving an overall range of 16-32 for the very worst and best candidates). Candidates at the ‘borderline’ (23.5 or 24 marks) are discussed in examination boards – and this again poses risks if candidate scores are adjusted (although we saw no evidence of this in the CR2B examination board which was very professionally conducted).

**Recommendation:** In conjunction with standard setting developments, it is feasible for the CR2B to design a system that requires a satisfactory ‘pass’ in each of the three components. Akin to the recommendations for the CO2B, we would encourage the oral to modify score sheets, with a series of high level domains, and overall global grade expressed in a letter/symbol rather than numeric format. This allows the College to consider the merits of weighting items, whilst hiding such weighting from examiners.

We would also recommend that examiners continue to mark separately, and only use the ‘wash up’ period to discuss if there is a significant difference within examiners pairs. Any such development should also ensure it retains the ability for examiners to provide narrative comments for feedback to candidates, which would be considered good practice.
FRCR Clinical Oncology

Overview

The FRCR (CO) is a well established assessment system that incorporates knowledge application, clinical short case and clinical oral examination formats. There are clearly defined entry routes into the CO assessment sequence, which comprises a balanced series of ‘modular’, knowledge based papers which are followed by a clinically oriented, integrated test of knowledge application. The final sequence comprises a clinical performance test involving direct patient encounters and a carefully structured oral examination.

Each component part will be discussed in greater detail, but some overarching themes within the CO assessment sequence merit commendation:

1) As noted, those assessment functions undertaken by the College are well described and function well. As the only examination that directly involves patients, the CO2B takes particular care in the recruitment and participation of patients, with scrupulous attention to detail, including comprehensive incident management planning
2) Clearly laid down curriculum outcomes that are available to all stakeholders
3) Development and delivery of written, knowledge test items using SBA formats. Accompanying this is a successful process of best practice question writing guidance, emendation, examiner support and accessibility of material by candidates
4) Feedback to unsuccessful candidates is well conceived and of high quality, including face-face delivery, counseling and engagement with educational supervisors for those candidates on UK Specialty Training schemes
5) Ongoing enhancements of the CO assessment process were obvious to the review team, in both written documentation and plans for development, and in delivery of an exceptionally well structured CO2B oral examination
There are a number of overarching areas for development across the whole assessment system, with consequent recommendations

1) Development of a ‘use argument’. Clarity about the purpose of test, the level of training of candidate and the application of results is an important mechanism that justifies the choice of assessment. Development of such statements will assist candidates and examiners in understanding the scope of each assessment component, and help relate it to other in-training assessments. Of note, such statements are often clearly laid out in examiner handbooks, with narrative justification of areas of focus (e.g. an appropriate emphasis of technical/safety skills over more generic attributes) appropriate to the context of the assessment – and these statements should be shared with candidates.

**Recommendation:** That the CO examination produces a clear statement of purpose for each of the component and makes this publically available to all stakeholders. This could be easily adopted and unified from existing provision to examiners. (P)

2) Test Standardisation and equating. Akin to the majority of major educational organizations undertaking multi-site assessment, the CO examination uses multiple different diets of papers to help with security. However, there is no mechanism to identify whether a paper is of a similar standard to others, other than broader performance data. Given the lack of blueprinting in elements of the CO assessment system, it does risk producing a paper that may well differ significantly in difficulty, but is not compensated for by approaches to standard setting.

Score linking is a process whereby different papers (e.g. for different diets) are constructed to the same content and statistical specifications. Equating involves using the raw scores from these different papers onto a common scale, and using small statistical adjustments to account for differences in the difficulty of each paper. This allows the deployment of multiple papers over time, with the ability to yield scores that can be sensibly interpreted. Deploying multiple different papers also increases test security.
**Recommendation**: A simple approach would be to nest a series of anchor questions across papers to track whether performance data is consistent across different diets of candidates, or if not, seeks to understand this (e.g. if there is evidence that the different sittings of the exam are undertaken by candidates of different seniority). More sophisticated approaches include longer term moves to using test equating (e.g. by application of Item Response (IRT) theory based mechanisms), a technique used by other large educational institutions.

3) **Blueprinting**. As outlined in the CR overview, blueprinting is an area of assessment practice that is often conflated with tagging/description of items within an item bank. It is clear that questions are very well tagged in CO1, and that the grid used in CO2A aligns system (x axis) and stages of the treatment journey (y axis) - thus noting pre-treatment (science and diagnostics), treatment and post treatment (including prognosis, follow up and palliation) plus an emergencies section (in miscellaneous) which map directly onto the CO curriculum. This could easily be translated into a test specific blueprint, and the review team also noted the work underway to 'reverse engineer' blueprints.

**Recommendation**: The review team encourages the CO assessment system to continue its development of test specific blueprinting. The CO2A grid could form the basis of both individual assessment blueprints, as well as an overarching format to examine the spread of content across the whole assessment system.

4) **Quality loop**. A key part of any assessment system is the question ‘how well are my assessments working?’ This question is partly answered through this review process as a significant weight of evidence has been gathered, and in aspects of the CO assessment system (e.g. CO1) where good, item-level analytic data is routinely provided, as well as an open access digest of exam performance.

**Recommendation**: Routine generation of psychometric quality reports for the CO examinations should be produced and reported, as per the suggestions in Dr Patterson’s overview. An extensive range of resources in support are listed in the bibliography.
FRCR Clinical Oncology Part 1 (CO1)

This is a four part, modularized assessment that directly maps to a major part of the CO curriculum. Each modular test comprises of 40-50 SBA questions, and whilst this can be taken in any order of modules, only four attempts are permitted to complete the whole CO1 sequence.

There is an established CO1 board which is responsible for question writing, setting of papers, examiner recruitment and training, setting standards and routine examination board functions. It is governed by the recently established overarching FEB.

The following comments should be read in conjunction with the good practice and areas for development/recommendation in the CR overview.

Areas working well

1) Test format, as delivered through SBA formats is appropriate, with excellent question writing guidance and examiner support. Banking and tagging of items is routinely established, with an active process underway to enhance the question bank.

2) An established process of standard setting, using the Angoff method is deployed. This is supported by a range of post-hoc analytics which allow exam board chairs to review/emend or discard items which continue to underperform psychometrically within the CO1.

3) Feedback to candidates is a significant strength of the CO1, with good levels of domain/themed feedback on areas of strength and weakness. It is anticipated that improvements in banking and blueprinting will help enhance this further.

4) At the whole-test level, the CO1 is to be commended for its production of annual report, which is publicly available, outlining areas of overall candidate areas of strength and weakness.
Psychometric Commentary

Full analyses and discussion of the Part 1 CO examinations held in Spring 2014 are provided in the separate ‘Statistical Report on Fellowship Examinations for Clinical Radiology and Clinical Oncology’ and associated annexes.

Part 1 CO comprises four papers each of 40 to 50 single best answer MCQs. Three of the four papers have similar characteristics with reasonable cohort mean scores, while the fourth paper, Physics, has a rather lower mean than the others. The standard deviations for all papers are very large (the largest of all Clinical Radiology and Clinical Oncology examinations). All of the papers have good to very good reliability, as judged from the high values obtained in Spring 2014 for Cronbach’s α (range 0.83 to 0.91), although this may be due to the large standard deviations, which can inflate values for α.

Failure rates are high for the Cancer Biology paper and, at 49%, very high for the Physics paper.

The Part 1 CO examinations are arranged as a series of module assessments which can be taken over a period of two years time, and presumably includes resits for failed modules. There may be a case for more synoptic assessment material to be included in this phase of training and development to ensure that knowledge is integrated and does not decay over time. This may be difficult to achieve if candidates do not take the assessments in a particular sequence, and this objective is in any case achieved by the Part 2A examination.

One Part 1 paper (for Cancer Biology and Radiobiology in October 2014) has been seen as part of this review. The paper is composed of 50 single best answer questions of a simple, knowledge recall format. The questions are reminiscent of true-false items except that there is one true and four false options. The questions would be improved if they routinely used the “which is the best/most likely/most appropriate” format for the question lead-in. It would be unfortunate if the same criticism applied to all papers. Recommendations are explored in the following section.
Areas for development/recommendations

There are very few specific areas for development of the CO1, over and above those recommendations detailed in the CO overview. As such, enhancements are suggested alongside each point:

1) **Question content.** Whilst the CO1 uses a recognized and appropriate test format, with good item writing guidance, the Review Team found that many of the questions tended to focus on factual recall, rather than enhanced clinical content and knowledge application testing.

   **Recommendation:** The review team recognizes that writing high quality assessment items for complex scientific material is challenging, but a range of advanced writing workshops exist to provide this support, or could be commissioned by the College as part of faculty development.

2) **Standard setting.** The review team notes the clear use of Angoff as a recognized, criterion set process. An articulation of the ‘just passing’ candidate that is reviewed at each meeting/teleconference will be an important resource for training new examiners, and for use as reference during the pre- and post discussion scoring elements.

   **Recommendation:** The CO1 develops an articulated version of the just passing candidate. Support could be provided by the CO2A examination board who have their own descriptions of the just passing candidate that are refreshed regularly.

3) **Feedback.** As noted, this is an area of very good practice for unsuccessful candidates. This is an area that can be automated (via easily available software) in respect of delivering thematic feedback for areas of strength and weakness – and be provided for successful candidates too.

   **Recommendation:** The CO1 considers the provision of routine feedback to all candidates using the format offered to those who are currently unsuccessful.
Final FRCR Clinical Oncology Part A (CO2A)

This is a highly integrated, synoptic test of 240 SBA items, undertaken as 2x120 item papers with a rest period for candidates between papers. Entry is contingent on success in the CO1 assessment, and there are a limited number of attempts to undertake the CO2A.

There is an established CO2A board which is responsible for question writing, examiner recruitment and training and routine examination board functions. Setting the CO2A paper and standards is shared between the chairs of the CO2A board and members of the CO2B board. It is governed by the recently established overarching FEB.

The following comments should be read in conjunction with the good practice and areas for development/recommendation in the CR overview. The review team noted an extensive range of good practice in the CO2A

Areas working well

1) Question format and construction use high quality SBA items. There is excellent structural support and guidance for new examiners, with clear processes for review and emendation. As ‘older’ items are released from the bank, they are made available to candidates, representing a larger pool of practice questions.

2) Tagging of items is a clearly established process, but within this banking system is a thoughtful grid format that allows identification of each question against a core learning outcome/part of the curriculum. This is arranged by system (x axis) and stages of the treatment journey (y axis) - thus pre-treatment (science and diagnostics), treatment and post treatment (including prognosis, follow up and palliation) plus an emergencies section (in miscellaneous). Accompanying this, there are clear decision rules to sample in a balanced way across the curriculum. As noted, this would easily allow the generation of routine test blueprints for CO2A, and a format that other aspects of the CO and CR could adopt.
3) Standard setting is undertaken using the Angoff procedure, and moreover carried out within a well described approach. Whilst 2 panels are set up, to ensure face-to-face meetings are accessible across the UK, items are shared across both panels. Although previous cohort data is not made available until after standard setting has taken place, care should be taken as routine viewing of old data risks moving standard setters away from the just passing candidate and more to the performance of previous cohorts). There is also a well articulated description of the just passing candidate that examiners indicate is very helpful.

4) Routine post-hoc psychometric data is made available to examination boards to help judge assessment quality.

5) Feedback to candidates is a significant strength of the CO2A, with good levels of domain/themed feedback on areas of strength and weakness. It is anticipated that improvements in banking and blueprinting will help enhance this further.

6) At the whole-test level, the CO2A is to be commended for its production of annual report, which is publicly available, outlining areas of overall candidate areas of strength and weakness.
Psychometric Commentary

Full analyses and discussion of the Part 2A CO examination held in spring 2014 is provided in the separate ‘Statistical Report on Fellowship Examinations for Clinical Radiology and Clinical Oncology’ and associated annexes.

The Part 2A examination in Clinical Oncology is a 240-item single best answer (SBA) assessment, split into two halves, both taken on the same day. It is presumed here that all candidates are presented with the same SBA questions.

The examination in Spring 2014 had a good reliability, with a Cronbach’s alpha of 0.86 and a small standard error of measurement at 2.83% and the examination performed well. The passing standard for that examination was 60.42%. Since the mean score was 60.7, it is not surprising that a very high proportion of the cohort (48%), failed this assessment. The review noted that the CO examinations do not routinely use the Hofstee method of standard setting, and that the apparent disparity in position between CO and CR examinations is under discussion at the joint Fellowship Board.

The Part 2A examination is both synoptic and robust and achieves gold standard reliability on a relative narrow range of candidates’ scores. The items are models of good SBA construction, with often rich clinical vignettes and clear lead-ins seeking the ‘best’ answer. These items should be a model for College examiners in CR as well as CO and the Part 1 CO questions would benefit from this style of questioning (although the level of clinical detail might be lower).

Areas for development/recommendations

There are no specific areas for development of the CO2A, over and above those recommendations detailed in the CO overview.
Final FRCR Clinical Oncology Part B (CO2B)

This is a two component assessment, with 5 structured clinical short cases (which uses ‘real’ patients to assess clinical signs) followed by a complex, structured oral assessment with 8 ‘cases’. Both formats are assessed by a pair of assessors in each case/oral, who mark independently with an overall ‘global’ mark awarded by each assessor/case to generate an overall score across both components of the assessment. There are clear non-conjunctive rules that require candidates to achieve a minimum standard in both clinical cases and orals, in addition to securing an overall passing score.

The clinical cases and oral assessments are undertaken separately, but within a short timeframe of each other. The clinical cases are held in an appropriate environment (e.g. hospital clinical skills centre), and only once clinical centre is used per exam diet. Patients are recruited from a mixture of settings (outpatients and inpatients), although dependent on availability and this can pose challenges for case specificity and matching across different diets of the examination. Inpatient selection is handled by senior trainees, guided by a framework that attempts to ensure sampling across key clinical areas. Cases are discussed, and case complexity determined by pre-examination discussions with the senior examiner and station examiners. This assessment was observed by Dr Patterson, who remarked on the well organized and carefully delivered examination format with good support staff provision and well engaged examiners, although evident examiner variation in terms of prompting was present.

The structured oral examinations could be best described as a thoughtful mix of a structured written examination, clinical activity (e.g. determining radiotherapy treatment fields) and oral examination. The examination is delivered in a series of sound proof rooms, with candidates rotating between two pairs of examiners, and covering 4 cases per examiner pair. All material is delivered as PowerPoint slide show, which takes the candidate through each case, with introductory clinical material and a series of questions and tasks highlighted on screen which focus around clinical decision making, prescribing and advanced, complex management. Examiners ask supplementary structured questions to test candidates’ use of clinical evidence, or to seek justification for responses, but are often passive during the oral.
Each clinical oral case goes through a series of emendation phases (including applying a standardized examination format) with a panel of examiners to help balance complexity. A simple running schema allows the examination board to ensure each candidate’s oral set samples widely and appropriately across the curriculum.

The oral assessment was attended by Drs Fuller and Patterson. There was an excellent briefing to candidates and examiners, as well as good examiner support during the orals. The examination proceeded smoothly with good academic and administrative support. Examiner behaviour and conduct was of a very high standard, with very little inappropriate prompting.

There is a dedicated examination board for the CO2B, which meets regularly and is also responsible for examiner training and recruitment, as well as executing traditional functions after assessment diets have completed. The review team also note and commend the significant amount of development activity that has clearly been underway to enhance the CO2B over recent years.

The following comments should be read in conjunction with the good practice and areas for development/recommendation in the CR overview.

Areas working well

1) Authenticity of assessments can often be sacrificed in the pursuit of standardization across examinations. However, the review team noted the thoughtful approach to ensuring patients were a central part of the clinical cases component, and this represents a real strength of the assessment.

2) The construction, emendation and final selection/sampling of structured cases provides considerable opportunity for blueprinting a balanced oral examination, and for ensuring examiner centred behaviours (such as prompting) are reduced.

3) Examiner guidance is clear and comprehensive with a well resourced handbook and training for examiners – built around clearly described anchor statements for global marking/grading.

4) Feedback is a notable strength of this assessment, and examiner marking schedules contain dedicated space for narrative comment.
Psychometric Commentary

Full analyses and discussion of the Part 2B CO examinations held in Spring 2014 are provided in the separate ‘Statistical Report on Fellowship Examinations for Clinical Radiology and Clinical Oncology’ and associated annexes.

There are two assessment components for Part 2B in Clinical Oncology. The first is a 5-station clinical examination, the second uses 8 case discussions with images and reports presented on-screen as a structured oral examination. Both elements are interesting in design. The clinical stations use five real patients, providing a highly authentic assessment but with the concomitant problems of case-matching, case specificity and standardization of assessment experience across all candidates. Considerable efforts are made to calibrate the patients and agree lines of questioning before the candidates are assessed, but even so, the reliability, as measured by Cronbach’s α, of this assessment (on the basis of only one set of analyses) is low at 0.63 when 13 ‘stations’ are combined and just 0.29 for the five clinical stations alone. Increasing the number of stations in an attempt to improve reliability is likely to be impractical (particularly for the clinical element). A possible solution is offered below.

Also the overall failure rate at Part 2B in Spring 2014 at 40.4% of the cohort is high.

The structured orals are interesting in that they are given as a PowerPoint presentation which contains both the case materials and the questions to be answered for the observing examiners. This format, where candidates can in fact read out the questions to the examiners before answering them looks to work very well with a minimum of prompting or additional questions by examiners. The cases are prepared as different ‘sets’ mostly prepared afresh for each diet. The effect of varying sets was not investigated for this report but the reliability, while still on the low side, was better than most of the College’s clinical assessments. The reliability might be improved further by moving from an notional 8-point marking scheme (i.e. up to four marks from each examiner) for each of the eight structured oral cases to more open marking as highlighted in the recommendations below.
The correlation between the scores of candidates taking the clinical and structured oral components of Part 2B is very high. This implies that the skills tested in the clinical examination are very similar to those tested in the structured orals. It would be acceptable therefore to combine the assessments for the purposes of statistical analysis. This is done, in any case to determine the results. Efforts should be maintained to improve the reliability of the clinical component, if at all possible.

In both clinical and structured oral examinations, final marks are compressed to a four point scale. The four point scheme used by each examiner for the overall case score in the structured orals equates to: fail, borderline, pass and good pass standards, which are terms already used by assessors. These global judgments could be combined with development of the grade anchors used in the structured orals into a domain-based scoring system could be used to standard set the pass mark for these assessments using a borderline regression technique. This would produce a move away from the current close marking and fixed passing standards to systems more in-line with current best practice.
Areas for development/recommendations

1) **Purpose of assessment.** The purpose of the assessment is well articulated within the CO2B examiner handbook, but is neither available to candidates, nor immediately apparent how the CO2B draws on knowledge/competencies previously examined in CO1 and CO2A, or how the examination clearly links with workplace training. Development in this area will help form clear benchmarks for candidates and trainees about the standard being tested. Trainees questioned the time pressure of the oral examination, but on discussion with examiners, they pointed to the authenticity of the format, relating this back to clinical MDT meetings, where many cases will be reviewed, and decisions made in a time-limited format.

**Recommendation:** Whilst it is easy to regard oral and ‘real’ clinical case examinations as inherently less desirable than structured/standardized formats, a clear use argument for the CO2B will contribute to the robustness and defensibility of the assessment, whilst preserving the essence of the joint clinical and oral examination.

2) **Blueprinting.** The review team notes the efforts underway to ‘reverse engineer’ blueprints for the CO2B, and existing frameworks to provide simple running schema for clinical and oral examinations. Whilst this is an overarching theme for the entire CO assessment system, there is a critical need for high quality blueprinting to map the entirety of clinical cases and structured orals onto one blueprint to ensure that candidates are given an appropriately sampled examination. At present, there is no clear mechanism to prevent candidates seeing cases across both assessment components with considerable overlap – or more importantly, potentially failing to be examined in a critical domain.

**Recommendation:** That the CO2B board produce an overarching blueprint for the entirety of the clinical and oral examination. Such a blueprint should also detail decision rules to help justify the number of clinical and oral cases.
3) **Clinical cases examination.** The clinical cases examination is a key part of the CO2B’s authenticity, and brings many strengths. The review team are conscious of the many challenges of delivering clinical assessments using ‘real’ (rather than simulated) patients, and would encourage the CO examinations to continue this format, but with further work to improve test robustness. A key element of this focuses on how cases are selected and mapped to the overall test blueprint. The review team was unclear about how case selection between centres (e.g. Mount Vernon vs. other centres used in rotation) was managed to ensure consistency or compensate for complexity in the absence of a blueprint. Discussion with trainees highlighted the authenticity of using clinical cases (and liked this aspect of the exam), but expressed concerns about the selection of cases between centres.

**Recommendation:** As part of blueprinting, consideration should be given to the selection of clinical cases to ensure a comparable selection and complexity is maintained across centres to ensure candidates are given a consistent assessment, irrespective of centre or examination diet. The review team is conscious that this will be a longer term development – particularly with small pools of patient volunteers and the provision of adequate reserves.

4) **Standard setting.** Currently, no recognized criterion based standard setting is applied to the CO2B. There is good practice in the design of anchor statements and training for what effectively could become a global scoring system. Although there are clear non-conjunctive rules (minimum number of adequate clinical cases and oral stations must be passed), the lack of formal standard setting reduces the validity of this assessment.

**Recommendation:** That the CO2B examinations adopt a recognized method of standard setting. This is in part dependent on both the sample size within each cohort and the construction of scoring/marking instruments (below). Borderline methods have the greatest evidence for successful use in performance assessments, but typically become less stable if the total candidate number/assessment diet drops below 50. Angoff methods are also recognized as suitable for setting standards for small cohort performance tests. A range of support resources are detailed in the bibliography.
5) **Scoring:** This currently takes the format of an extended narrative with a global score based on anchor statements. The global mark is expressed numerically, which will have an impact on examiner behaviour – and global anchors are better expressed as letters. The score sheets could be modified to provide individual case/station based marksheets, each with a small series of pre-determined high level domains to score and an overall global grade. High level domains could include reasoning, planning, clinical management, use of evidence and be directly mapped against the blueprint. This may well require an extension of time for candidates to move between stations.

**Recommendation:** As part of moves towards formalized standard setting, the review team would encourage pilot work to design and test new forms of scoring sheets, looking carefully at cognitive load on examiners. Any such development should also ensure it retains the ability for examiners to provide narrative comments for feedback to candidates, which is good practice. Development of the scoring formats also presents a key opportunity to ensure the ‘added value’ of using real patients (rather than standardized) is captured in examiner scores and judgements.

6) **Post-hoc routine analysis.** As noted in Dr Patterson’s supplementary reports, the clinical cases alone do not provide an adequate spread of assessment episodes to provide a reliable test. An option could be to increase the number of clinical cases, although this is likely to pose significant logistical challenges. On deeper analysis, we note that there is a very strong correlation between candidate performance in the clinical and structured oral components of CO2B, suggestive that both examinations are testing similar/overlapping traits. Routine analysis should also measure any potential site effect (for the clinical), time effects or differences due to examiner pairings. Such analyses provide powerful validity evidence, and can form the basis of an ongoing quality improvement loop.

**Recommendation:** We would suggest a common scoring system and that both clinical cases and structured orals are combined for the purposes of overall reliability analyses. Irrespective of format of standard setting (Angoff or Borderline methods), simple post hoc analysis should be routinely performed.
Equality and Diversity/Protected Characteristics analysis

A number of factors have hindered a full analysis of the ‘protected characteristics’ of CR and CO Fellowship candidates. Factors include: lack of availability of sufficient data in a convenient form to permit merging of candidate performance data across several iterations of the examinations (to increase sample size and to average test to test variations); the (limited) extent to which candidates have disclosed some of the protected characteristics to the College; the technical difficulty of performing sufficiently rigorous analyses of interactions between cohort sub-groupings.

Since the area is of considerable interest and potential concern to the College, some limited analyses have nevertheless been performed for this review. The small numbers of candidates taking CO examinations in Spring 2014 mean that probably no meaningful analysis at all can be performed at this stage, but analyses have been attempted for the Part 1 and Part 2 CR examinations held at the same time where candidate numbers are greater.

It must be stressed that the techniques used here are of the simplest, selected in a spirit of parsimony (One-way ANOVA, with some additional tests of significant differences), and that analyses are confined only to first attempt candidates because of the limited data. It is essential to note that there are multiple confounding factors. The cohorts are sub-divided by gender, place of primary medical qualification or ethnicity for each test, so, for example, effects of ethnicity may confound conclusions drawn about gender. Also some examinations are run as different ‘sets’, while for others, such as CR2A, different groups of candidates may take each Module assessment. The data provided by the College are therefore rather messy and results from analyses should be interpreted with caution.

Analysis not only requires larger samples (many of the ethnic groups fail the ‘Rule of ten’ for inclusion in any analysis) but may requires the use of more refined statistical techniques such as multiple linear regression to control for the many variables. Given these qualifications, the following is a reasonable summary of major associations between some protected characteristics and candidates’ scores across one diet of Part 1 and Part 2 examinations in Clinical Radiology:
**Gender:** Female average scores are sometimes higher than male scores, but the differences in mean scores are typically small (<2%) and the differences are not statistically significant at the 5% level across Parts 1 and 2. The exception is CR1 Physics where females scored statistically significantly lower marks than males. It is of note that CR1 uses true/false questions, an assessment format where there is evidence that males typically obtain high scores.

**Place of primary medical qualification:** The cohort can be divided in UK, EU and Overseas (OS) candidates on the basis of country of nationality, although this should not assume country of primary medical qualification. The number of EU candidates is small and where they are included in the analyses, they are not significantly different from the UK group. As a single group, overseas candidates obtain statistically significantly lower scores than UK candidates in all CR Part 1 and Part 2 examinations, other than two modules in CR2A where OS mean scores are not significantly different from UK scores.

**Ethnicity** data is much more difficult to evaluate since the numbers of candidates can be very small in some ethnic groupings. At Part 1 CR Anatomy, candidates classified as Asian (ASI) look to have performed significantly less well than many other groups, but numbers in all groups are small. In Part 1 Physics, the lowest mean scores were obtained by Asian (ASI), Asian Other (AOTH) and Indian (IND) candidates and the mean scores were significantly lower (p≤0.001) than the White British (WHB) group.

At Part 2A the picture is more obscure because there are some significant differences between scores on the 6 papers, which is a major confounder for comparing overall performance by different ethnicity groupings. Whilst nine ethnic groups meet the ‘Rule of 10’ where scores are polled across the 6 part 2A modules, only Indian (IND) and White British (WHB) meet the ‘Rule of 10’ for individual modules, and Chinese (CHI) has been included in the more detailed analysis because it only just fails to meet the criterion. Using the Bonferroni correction in significance tests on these remaining groups suggests that no significant differences exist between IND, WHB and CHI in individual modules scores and pooled 2A module cores.
The situation is considerably more complex at CR2B where only IND and Pakistan (PAK) meet the ‘Rule of 10’. This is presumably because very few candidates provided ethnicity information, although 110 candidates took the examination. Comparing the performance of 25 IND with 11 PAK candidates showed no significant difference in scores for any of the three elements of CR2B

**Summary**

Better governance and data acquisition/handling policies should ensure the College is able to better examine key questions about the impact of ethnicity, training environment/location and place of primary medical qualification on achieving the FRCR. Such work should also examine progression within the FRCR examination suite, and could conceivably be undertaken as a partnership with interested external organizations to proactively research these matters. As noted in the main body of the report, the lack of test equating/anchoring across assessments means that it is possible that additional confounders such as item sampling and standard setting for OS examinations are contributing. With better (and larger) data sets, linear regression techniques will be able to control for such confounders and re-examine in difference in performance.

This summary on protected characteristics should be read alongside the detailed narrative commentary in this report and the ‘Statistical Report on Fellowship Examinations for Clinical Radiology and Clinical Oncology’ and associated annexes.
Summary

The final component of this review summarises a number of high level areas working well, and further recommendations/areas for development against the five areas of the GMC’s Standards guidance. This summary should be read alongside the detailed narrative reporting in this report, and the ‘Statistical Report on Fellowship Examinations for Clinical Radiology and Clinical Oncology’.

Each assessment is summarized in turn, beginning with CR1 overleaf:
<table>
<thead>
<tr>
<th>CR1</th>
<th><strong>Standards guidance</strong></th>
<th><strong>Areas working well</strong></th>
<th><strong>Areas for development</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Planning</strong></td>
<td>Good externally support facing material for candidates and examiners, including opportunity to undertake practice questions for anatomy and physics components</td>
<td>Development of a clear ‘purpose of assessment’ statement as part of wider enhancements within FRCR CR</td>
<td>Specific focus on test format for Physics component – move from T/F questions to SBA, and articulate if knowledge, or knowledge application test</td>
</tr>
<tr>
<td><strong>Content</strong></td>
<td>Examination authenticity and radiological anatomy assessment.</td>
<td>Requirement to deliver prospective examination blueprints for both components</td>
<td></td>
</tr>
<tr>
<td><strong>Delivery</strong></td>
<td>Clear and commendable College processes for assessment delivery and review</td>
<td>Appropriate use of post hoc analysis to generate appropriate routine item level measures that are pertinent to the test format</td>
<td>Routine generation of KR/reliability data from highly reliable examinations</td>
</tr>
<tr>
<td><strong>Outcomes</strong></td>
<td>Good use of Angoff as standard setting mechanism, with well described process to support examiners and standard setters.</td>
<td>Written articulation of the just passing candidate to assist new examiners, question writers and standard setters</td>
<td>Currently no provision of feedback to candidates. Requirement to generate routine data to support candidates (whether pass or fail)</td>
</tr>
<tr>
<td><strong>Review</strong></td>
<td>Excellent assessment support structures delivered by College (organization, security, governance and SOPs). Noteworthy major Automation project to come online from 2016 to enhance assessment resource and processes.</td>
<td>Pan CR and CO recommendations at College level – strengthening remit of FEB, documentary and data governance enhancements</td>
<td></td>
</tr>
</tbody>
</table>
### CR2A

<table>
<thead>
<tr>
<th>Standards guidance</th>
<th>Areas working well</th>
<th>Areas for development</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planning</td>
<td>Good externally support facing material for candidates and examiners, including release of old questions to candidates. Use of established, recognised test format (SBA)</td>
<td>Development of a clear 'purpose of assessment' statement as part of wider enhancements within FRCR CR. Specific focus on rationale of modular test format within assessment purpose statement.</td>
</tr>
<tr>
<td>Content</td>
<td>Comprehensive item tagging/mapping.</td>
<td>Requirement to deliver prospective examination blueprints.</td>
</tr>
<tr>
<td>Delivery</td>
<td>Clear and commendable College processes for assessment delivery and review. Reasonable reliability measures for each modular assessment.</td>
<td>High failure rates merit investigation – candidate preparedness and adequacy of standard setting judgments are key factors for consideration.</td>
</tr>
<tr>
<td>Outcomes</td>
<td>Recognised use of Angoff as standard setting mechanism.</td>
<td>Currently no provision of feedback to candidates. Requirement to generate routine data to support candidates (whether pass or fail).</td>
</tr>
<tr>
<td>Review</td>
<td>Excellent assessment support structures delivered by College (organization, security, governance and SOPs).</td>
<td>Pan CR and CO recommendations at College level – strengthening remit of FEB, documentary and data governance enhancements.</td>
</tr>
</tbody>
</table>

### CR2B
<table>
<thead>
<tr>
<th><strong>Standards guidance</strong></th>
<th><strong>Areas working well</strong></th>
<th><strong>Areas for development</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Planning</strong></td>
<td>Clearly articulated tri-partite assessment system with considerable authenticity and strong clinically oriented focus</td>
<td>As part of wider CR review commentary, articulation of an externally facing 'purpose of assessment' statement</td>
</tr>
<tr>
<td></td>
<td>Good externally support facing material for candidates and examiners with highly commendable examiner guidance</td>
<td>Review of current oral examination format is urgently required to maintain CR2B currency. Scope to adopt similar model to CO2B whilst preserving oral format</td>
</tr>
<tr>
<td><strong>Content</strong></td>
<td>Examination authenticity – particularly in respect of Rapid Reporting and aspects of oral examinations.</td>
<td>Development of prospective blueprinting to map entirety of reporting and oral cases to ensure an appropriately sampled, broad scope assessment. More standardization of case selection across candidates desirable.</td>
</tr>
<tr>
<td><strong>Delivery</strong></td>
<td>Clear and commendable College processes for assessment delivery and review. Excellent briefing for examiners and candidates.</td>
<td>Currently no format to undertake detailed item level psychometric analysis – and low reliability of CR2B elements. Suggested emendation to scoring and standard setting mechanisms should permit generation of whole-test and item level data for analysis, and improve reliability.</td>
</tr>
<tr>
<td><strong>Outcomes</strong></td>
<td>Feedback to unsuccessful candidates is comprehensive</td>
<td>Adoption of recognised criterion based standard setting for oral (e.g. Borderline or Angoff) dependant on cohort sizes. Separate criterion referenced standard setting for Reporting elements required.</td>
</tr>
<tr>
<td><strong>Review</strong></td>
<td>Excellent assessment support structures delivered by College (organization, security, governance and SOPs). Noteworthy major Automation project to come online from 2016 to enhance assessment resource and processes.</td>
<td>Pan CR and CO recommendations at College level – strengthening remit of FEB, documentary and data governance enhancements</td>
</tr>
<tr>
<td><strong>CO1</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Standards guidance</strong></td>
<td><strong>Areas working well</strong></td>
<td><strong>Areas for development</strong></td>
</tr>
<tr>
<td>------------------------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
</tbody>
</table>
| **Planning**           | Good externally support facing material for candidates and examiners, including release of old questions to candidates  
                       | Use of established, recognised test format (SBA). Strong governance and reporting, including annualized assessment ‘digest’  
                       | Development of a clear ‘purpose of assessment’ statement as part of wider enhancements within FRCR CO  
                       | Continued development of item bank to focus on knowledge application testing and enhanced clinical relevance                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |
| **Content**            | Comprehensive item tagging/mapping and active enhancement of item bank  
                       | Routine emendation and review of items.  
                       | Routine prospective blueprinting of modular components to overarching framework (scope to adapt CO2A blueprint)                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| **Delivery**           | Clear and commendable College processes for assessment delivery and review  
                       | Routine generation of post-hoc analysis and item review. Modular components are highly reliable  
                       | High failure rates merit investigation – candidate preparedness and adequacy of standard setting judgments are key factors for consideration.                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
| **Outcomes**           | Good use of Angoff as standard setting mechanism, with well described process to support examiners and standard setters.  
                       | Feedback to unsuccessful candidates is comprehensive  
                       | Consider offering routine feedback to all candidates (not just those who have failed)  
                       | Written articulation of the just passing candidate to assist new examiners, question writers and standard setters. Scope to adapt processes used by CO2A                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
| **Review**             | Excellent assessment support structures delivered by College (organization, security, governance and SOPs).  
<pre><code>                   |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | Pan CR and CO recommendations at College level – strengthening remit of FEB, documentary and data governance enhancements                                                                                                                                                                                                                                                                                                                                                                                                                  |
</code></pre>
<table>
<thead>
<tr>
<th>Standards guidance</th>
<th>Areas working well</th>
<th>Areas for development</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planning</td>
<td>Good externally support facing material for candidates and examiners, including release of old questions to candidates</td>
<td>Development of a clear ‘purpose of assessment’ statement as part of wider enhancements within FRCR CO</td>
</tr>
<tr>
<td></td>
<td>Use of established, recognised test format (SBA). Strong governance and reporting, including annualized assessment ‘digest’</td>
<td></td>
</tr>
<tr>
<td>Content</td>
<td>Generation of high quality test items which test knowledge application</td>
<td>Consideration of more routine use of ‘anchor’ items across papers to examine stability and cohort attainment</td>
</tr>
<tr>
<td></td>
<td>A comprehensive system of item mapping/tagging with a ‘patient journey’ type blueprint that is linked to the curriculum – ensuring appropriate and broad sampling</td>
<td></td>
</tr>
<tr>
<td>Delivery</td>
<td>Well delivered, highly reliable assessment. Routine generation of post-hoc analysis and psychometric review of items is associated with emendation and review of items</td>
<td></td>
</tr>
<tr>
<td>Outcomes</td>
<td>Good use of Angoff as standard setting mechanism, with well described process to support examiners and standard setters.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Feedback to candidates</td>
<td></td>
</tr>
<tr>
<td>Review</td>
<td>Excellent assessment support structures delivered by College (organization, security, governance and SOPs).</td>
<td>Pan CR and CO recommendations at College level – strengthening remit of FEB, documentary and data governance enhancements</td>
</tr>
</tbody>
</table>
### CO2B

<table>
<thead>
<tr>
<th>Standards guidance</th>
<th>Areas working well</th>
<th>Areas for development</th>
</tr>
</thead>
</table>
| **Planning**       | Clearly articulated bi-partite assessment system with considerable authenticity and strong clinically oriented focus.  
  Development of structured oral with emendation and routine review of cases  
  Clear, comprehensive Examiner guidance | Development of a clear ‘purpose of assessment’ statement as part of wider enhancements within FRCR CO.  
  Specific focus for the CO2B to justify time allowances within stations and structured orals |
| **Content**        | Authentic clinical case material and highly structured oral cases which are mapped to key learning components  
  Active development of blueprinting (e.g. current ‘reverse engineering’) | Development of prospective blueprinting to map entirety of clinical and oral cases to ensure an appropriately sampled, broad scope assessment. |
| **Delivery**       | Clear and commendable College processes for assessment delivery and review | Currently no format to undertake detailed item level psychometric analysis – routine reliability measures are feasible to undertake (combining clinical and oral cases) and suggested emendation to scoring and standard setting mechanisms should permit generation of whole-test and item level data for analysis |
| **Outcomes**       | Feedback to candidates, including dedicated score sheet design to gather examiner narratives | Adoption of recognised criterion based standard setting (e.g. Borderline or Angoff) dependent on cohort sizes. |
| **Review**         | Excellent assessment support structures delivered by College (organization, security, governance and SOPs). | Pan CR and CO recommendations at College level – strengthening remit of FEB, documentary and data governance enhancements |
Appendices and bibliography

Review team biographical information

**Dr Richard Fuller**

Richard Fuller graduated MBchB from Leeds School of Medicine in 1995. He returned to read his higher degree in Leeds whilst undertaking higher specialist training in Geriatric Medicine in Yorkshire. His current role is split between clinical work as a Consultant Geriatrician and as Director of Undergraduate Education at Leeds Institute of Medical Education. He is responsible for directing the MBChB degree programme which has recently undergone a significant and successful curriculum review.

His main research interests focus on assessment, particularly in relation to assessment frameworks and the quality measurement and improvement of criterion based assessment, publishing regularly with colleagues in Leeds and presenting at leading education conferences. His current work focuses on OSCE development, assessor behaviour, workplace assessment and the use of mobile technology to facilitate assessment development. He advises institutions nationally and internationally on both curriculum design and assessment matters, acting as a programme reviewer, external examiner and specialist advisor to a number of Schools. He has undertaken assessment work for the GMC, Royal College of Physicians and National School of Healthcare Sciences. He is a GMC Education Associate, leading the National Review of Assessment in Undergraduate Medical Schools.

**Dr John Patterson**

Until retiring from full-time employment in September 2009, John, a physiologist, was Associate Dean for Undergraduate Medical Studies and Head of MBBS Assessment at Barts and the London School of Medicine and Dentistry. During a long career at Barts he became interested in medical education, eventually in 1998 becoming Head of Undergraduate Medical Studies. His educational interests include: curriculum development; assessment design and analysis; degree regulations and quality assurance. In retirement, John gives talks and runs workshops on good assessment design, assessment metrics and standard setting techniques for the health care professions. He also acts as an assessment consultant to various higher education institutions, Royal Colleges and regulatory bodies.
His recent published work includes a collection of single best answer and extended matching questions complete with feedback in, Jade Chow and John Patterson (2012), ‘Oxford Assess and Progress in Medical Sciences’ Oxford University Press. 480pp. He has contributed regularly to the Association for Medical Education in Europe with conference papers on standard setting, progress testing and OSCE examinations.

**Professor Trudie Roberts**

Professor Roberts graduated from Manchester with a degree in Medicine and a BSc in Anatomy. She undertook her early medical training in Manchester and her research in Manchester and the Karolinska Institute in Sweden. In 1995 she was appointed Senior Lecturer in Transplant Immunology at the University of Manchester. In 2000 she was appointed Professor of Medical Education at the University of Leeds. She was awarded a National Teaching Fellowship in 2006. In 2009 she was appointed Director of the Leeds Institute of Medical Education.

She was a council member of the General Medical Council from 2009 until 2012 and Chair of the Association for the Study of Medical Education until July 2013. She is a council member and Censor for the Royal College of Physicians of London. In September 2013 she took over as President of the Association for Medical Education in Europe. Professor Roberts's main interests and expertise are in the areas of assessment of competence, professionalism, and transitions in training and education.

**Professor Katharine Boursicot**

Kathy graduated from the Medical College of St Bartholomew's Hospital in London, with a degree in Medicine and a BSc in Anatomy, and then spent 18 years as an Obstetrician Gynaecologist in London, Dublin and Hong Kong.

Kathy has been the Specialist Assessment Advisor to the Society of Apothecaries Examinations Committee for 6 years. After completing a Masters in Education, Kathy moved full-time into medical education. She has worked at Barts and The London School of Medicine and Dentistry, Hong Kong University Faculty of Medicine, the Clinical School of Medicine at the University of Cambridge and St George's University of London, before taking up the post of Head of Assessment at Lee Kong Chian School of Medicine in Singapore.
Kathy has acted as a consultant to the GMC on the development of tests of competence for Fitness to Practise Procedures and the PLAB Part 2 examination. She was the Treasurer of the Association for the Study of Medical Education (ASME) and chaired the Board of Management of the journal Medical Education. She has been a Specialist Advisor to the Higher Education Academy’s Subject Centre for Medicine, Dentistry and Veterinary Medicine (MEDEV) for 8 years. Her main research interests are in the fields of assessment of clinical competence and professionalism, standard setting and workplace-based assessment; she has published on standard setting, OSCEs, equivalence of standards and diversity issues in medicine.

With 18 years’ experience of designing and implementing OSCEs at undergraduate and post-graduate levels in medicine, dentistry and veterinary medicine, Kathy has been invited to advise on the conduct of assessment, especially OSCEs, in the UK, Denmark, Jordan, Malaysia, Myanmar, Norway, Qatar, Saudi Arabia, Singapore, South Africa, Sudan and Sweden. She led the Theme Group on Assessment of Clinical Performance at the Ottawa International Conference on Assessment in 2010, which reviewed the literature for best practice in OSCEs and WBAs. She also has 16 years' experience of standard setting for written tests & OSCEs.
References and support resources

A range of comprehensive assessment resources (including assessment training workshops and conferences) are detailed below, with specific resources linked to areas of recommendation.

Assessment workshops and conferences

- Fundamentals in Assessment Course: [http://facourse.webs.com](http://facourse.webs.com)
- Essentials in Medical Education Assessment (AMEE conferences):
- Fundamentals of Assessment in Medical Education (Ottawa conferences):
- International Advanced Assessment Course: [http://www.iaacourse.com](http://www.iaacourse.com)

Validity and Standards in Educational Assessment and Testing

- Kane, MT. Validating interpretive arguments for licensure and certification examinations. Evaluation and the Health Professions. 1994;17(2):133-159
- Downing, SM. Validity: on meaningful interpretation of assessment dataMedical Educ. 2003;37(9):830-837
Blueprinting

- General Medical Council PLAB blueprint: [http://www.gmc-uk.org/doctors/plab/Blueprint.asp](http://www.gmc-uk.org/doctors/plab/Blueprint.asp)
- Sales et al. Blueprinting for clinical performance deficiencies – Lessons and principles from the General Medical Council’s fitness to practice procedures. 2010;32(3):e111-114

Standard Setting

- Standard setting is an area that is best undertaken as a practical exercise – all of the assessment workshops described above run a series of practical components that focus on Angoff and Borderline methods

Psychometrics and quality measures
