

**Royal College of Radiologists Assessment Review:**

**Statistical Report on Fellowship Examinations for Clinical  
Radiology and Clinical Oncology.**

**Prepared by:  
John Patterson, PhD  
December 2014  
(Final version)**

## CONTENTS

Section	Title	Page
1	Summary	3
2	Introduction	8
3. Fellowship of the Royal College of Radiologists in Clinical Radiology		
3.1	Part 1 Examinations held in Spring 2014.	8
3.2.	Part 2 Examinations held in Spring 2014.	
3.2.a	Final Part 2A examinations	11
3.2.b	Final Part 2B examinations	12
4 Fellowship of the Royal College of Radiologists in Clinical Oncology		
4.1	Part 1 Examinations held in Spring 2014.	17
4.2	Part 2 Examinations held in Spring 2014.	
4.2.a	Final Part 2A examinations	19
4.2.b	Final Part 2B examinations	20
Available separately		
Annex 1A	Fellowship of the Royal College of Radiologists in Clinical Radiology: Whole-test analyses for Spring 2014.	
Annex 1B	Fellowship of the Royal College of Radiologists in Clinical Radiology: Item-level statistics for Spring 2014.	
Annex 2A	Fellowship of the Royal College of Radiologists in Clinical Oncology: Whole-test analyses for Spring 2014.	
Annex 2B	Fellowship of the Royal College of Radiologists in Clinical Oncology: Item-level statistics for Spring 2014.	
Annex 3	Glossary of psychometric and statistical terms.	
Annex 4	Analyses of protected characteristics for Clinical Radiology. -- <i>to follow</i>	

## Summary

This executive summary provides some general conclusions and recommendations based mainly on the statistical analyses of the Spring 2014 examinations for Parts 1 and 2 of the Fellowships in Clinical Radiology and Clinical Oncology. The full report, which follows, gives more detail, and statistical summaries are provided in separate annexes. The analyses and report were undertaken as part of an external review of these examinations. A critique of documentary evidence supplied to the reviewers is the subject of a separate report. Some comments about the general administration and conduct of the examinations are included here following observation of the Part 2B clinical examinations held in October 2014.

### General:

The general administration of the examinations is exemplary and we have been impressed by the thought and care that goes into assessment delivery in terms of venues, briefing of candidates and examiner training. With a few exceptions the CR and CO assessments perform to an acceptable or good standard. For the tests of knowledge and its application, the assessments generally perform well, with acceptable to good reliability of scores. Item design has not been studied in detail for this report although sample papers for the October 2014 diet have been examined and visits were made to observe the Part 2B examinations for both Fellowships. The clinical assessments in parts 2B are generally well-conceived, but, as is often the case when testing clinical skills, reliability is often lower than desirable for a high-stakes assessment.

### Recommendations

#### A) College administration

While College administration performs well in terms of documentation and published procedures, some benefits would be obtained from the following:

:

**A1)** There is a clear need to establish better data-basing of candidate personal (and protected) characteristics. As a minimum this should contain the performance (with final marks as a minimum) for each candidate's assessment history while taking College examinations. This should also give the re-sit history and re-sit scores for every candidate. Such a database will allow longitudinal studies of candidates across their various examinations and will in time generate sufficiently large pools of candidates to be able to compare the performance of candidates by protected characteristic, re-sit attempts, gender and place of primary medical qualification and training, etc.

**A2)** Analysis of protected characteristics is hampered by many candidates not offering the appropriate information. While candidates cannot be compelled to do so, steps should be taken to optimise collection of this information, since there is a general expectation that Colleges will wish to monitor for any potential bias in their examinations.

**A3)** In addition to A1) above, a formalised approach should be adopted for the recording of candidates' scores on every item and their overall score for every examination. There must also be clear linkage between the candidate number for each examination and their unique identifier in the collective database.

**A4)** At present, basic examination statistics are provided for each 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.

## **B) The Examinations in Clinical Radiology**

**B1)** The examinations in Anatomy in Part 1 are given as four 'sets', which are of the 'name the structure' format and are hand-marked. These sets differ in content and probably also in item difficulty. While it is 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<sup>1</sup>. The tests perform well in statistical terms.

**B2)** The Part 1 Physics examination also performs well, with high reliability. It does however employ true-false MCQ items which are no longer considered the best MCQ format. Such items have been surpassed by the SBA and EMQ formats. It is hard to recommend change to what looks to be an effective assessment, but a switch to SBAs should be seriously considered. The true/false MCQs in Physics also look to be testing solely factual recall. Other Part 1 papers have not been examined to see if they test reasoning skills and the application of knowledge. Such testing of cognitive skills would be appropriate, even at Part 1, in postgraduate assessments.

**B3)** The Part 2A examination comprises six, module-based assessments which candidates can sit over a period of time. While this 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.

**B4)** The Part 2A best of five papers, to judge from that for Module 1 (Cardiothoracic and Vascular) for October 2014, use 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. This construction should be avoided.

**B5)** 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.

---

<sup>1</sup> This may already be the case, but no standard setting documentation (other than pass marks) has been supplied for these examinations.

While this situation is probably unavoidable, the points mentioned in B1 concerning Part 1 Anatomy should be borne in mind. The third element comprises two oral examinations where a (sometimes variable) number of cases are discussed. The reliability (Cronbach's  $\alpha$ ) for Part 2B overall and for each element 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. More marks for each item might improve the discrimination of the test and this in turn should improve reliability. For both reports and orals, the marking scheme could be widened and final marks not compressed as in the current scale. 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.

### **C) The Examinations in Clinical Oncology**

**C1)** The Part 1 examinations are arranged as a series of module assessments which can be taken over a variable period of time. 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.

**C2)** One Part 1 paper, that for Cancer Biology and Radiobiology in October 2014, has been seen as part of this review. The paper is composed of 60 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 is also unfortunate that few if any questions provide clinical scenarios for what is a postgraduate medical examination and it would be a pity if similar criticisms applied to all four Part 1 papers. Nevertheless the four Part 1 papers perform well in statistical terms and look to have good to very high reliability coefficients. The latter probably arises from the very wide spread of marks scored.

**C3)** The Part 2A examination is both synoptic and robust. It comprises 240 single best answer items and achieves gold standard reliability on a relative narrow range of 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).

**C4)** The Part 2B examination is in two parts, a 5-station, clinical examination and eight, case-based, structured oral questions. 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 standardisation 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 of this assessment is (on the basis of only one set of analyses) unacceptably low. Increasing the number of stations in an attempt to improve reliability is likely to be impractical. A possible solution is offered in C6, below.

**C5)** The structured orals are interesting in that they are given as a PowerPoint presentation which is operated by the candidate and which contains the questions to be answered for the observing examiners. This format looks to work very well with a minimum of prompting or additional questions by examiners. The cases are prepared as different 'sets' drawn from a bank of cases. 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 8-point marking scheme to more open marking.

**C6)** The correlation between the scores of candidates taking the clinical and structured oral components of Part 2B is weakly positive. This implies that the skills tested in the clinical examination are similar but not identical to those tested in the structured orals. It might nevertheless be acceptable 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.

**C7)** In both clinical and structured oral examinations, final marks are compressed to a four point scale. This could be equated to: fail, borderline, pass and good pass standards. These global judgements combined with development of the grade anchors used in the structured orals into domain-based scores 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.

#### **D) Protected characteristics.**

Unfortunately, a number of factors have hampered a full analysis of the so-called protected characteristics. Factors include: 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 sometimes limited extent to which candidates choose to disclose some of the protected characteristics to the College; the technical difficulty of performing sufficiently rigorous analyses of interactions between cohort sub-groupings. 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 since candidate numbers are greater for CR than for CO..

It must be stressed that the techniques used here are of the simplest. One-way ANOVA has been used, and data has been assumed to be of equal variances, although it is not always safe to make this assumption. In addition, there are multiple confounding factors. The cohorts are sub-divided by gender, place of primary medical qualification or ethnicity for each test, as, 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 are therefore rather messy and the results of analyses should be interpreted with caution. Proper analysis not only requires larger samples (many of the ethnic groups fail the 'Rule of ten' for inclusion in any analysis) but may require 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 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 Part 1 Physics where females scored statistically significantly lower marks than males. It is interesting that this assessment uses true-false MCQs where there is evidence that males typically obtain higher scores.

**Place of primary medical qualification:** The cohort can be divided into UK, EU and Overseas (OS) candidates on the basis of nationality and therefore, presumed 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 CR examinations except for two modules in Part 2A 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) 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 of these groups were all significantly lower ( $P \leq 0.001$ ) than 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. While nine ethnic groups meet the 'Rule of 10' when scores are pooled across the 6 Part 2A module, only Indian (IND) and White British (WHB) meet the 'Rule of 10' for the individual modules, and Chinese has been included in 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 both individual module scores and pooled module scores. The situation is even worse at CR2B where only IND and PAK meet the 'Rule of 10' requirements. This is presumably because very few candidates provided the ethnicity information, although some 110 took the examination. Comparing the performance of the 25 IND with the 11 PAK candidates showed no significant difference in scores for any of the three elements of CR Part 2B.

**Other factors:** All of the above analyses were performed on **first attempt candidates only** in the Spring 2014 diets of examinations. One factor that should be analysed when sufficient numbers are available is the retake history of candidates and the relationship between failure at the first attempt and the likelihood of passing at subsequent attempts. This might

provide some evidence to guide policy on the number of attempts and time frame for resit candidates.

## **Introduction**

This report summarises the findings of statistical analyses of candidates' scores in the Parts 1 and 2 examinations for the FRCR in Clinical Radiology (CR) and Clinical Oncology (CO). These analyses provide a health check on the outcomes of the FRCR assessments and are a part of the current (2014) Review of RCR examination procedures.

Almost all the analyses used here are taken from classical test theory. Thus far only results from the Spring 2014 diets of examinations have been analysed. While it is of interest to examine correlations between Parts 1 and 2, the varying time scales by which candidates achieve the final award and the different item contents of the various papers pose problems in terms both data extraction and data interpretation. Thus an examination of predictive validity – the extent to which Part 1 scores predict Part 2 scores - is probably outside of the scope of the present review. Similarly, because at Part 2 candidates may not take Part 2A and Part 2B at the same sitting, it may not be possible to examine correlations between these Parts to study concurrent validity. With only 17 candidates sitting both Part 2A and Part 2B in Clinical Oncology in Spring 2014, such analysis may not be useful. It has been attempted for the 33 candidates (out of 397) who sat all elements of the Part 2 Clinical Radiology examinations in Spring 2014 but obviously these candidates do not represent the norm and the results should be interpreted with caution.

This report is in sections, each describing the performance of one part of either CR or CO examinations. Fuller summaries of the statistics for each paper are found in Annexes 1A and 1B for Clinical Radiology and Annexes 2A and 2B for Clinical Oncology. For example, whole-test statistics for CR are provided in Annex 1A, while item-level statistics for the individual CR papers can be found in Annex 1B.

## **Fellowship of the Royal College of Radiologists in Clinical Radiology**

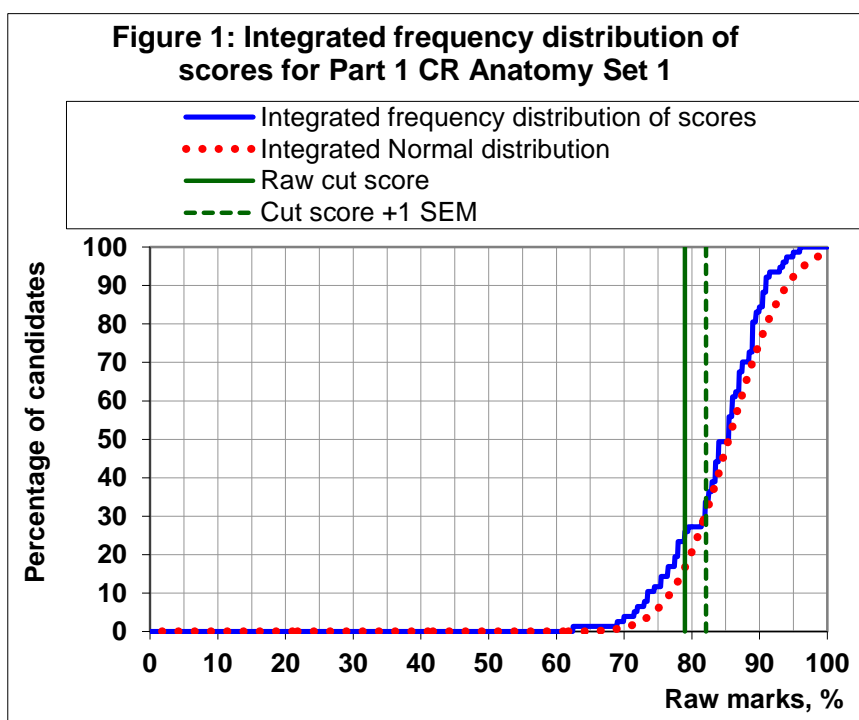
### **3.1 Part 1 Examinations held in Spring 2014.**

Table 1 shows the basic, whole-test statistics for the Part 1 examinations in Anatomy and Physics. The Anatomy questions are of the 'name the structure' format and are hand-marked. The Physics questions are groups of five true/false items per stem (without negative marking). Taken as a group the five sets of papers in Part 1 CR have broadly similar characteristics with fairly high cohort mean scores and, in most cases, large standard deviations. All of the papers have acceptable to very good reliability, as judged from the high values obtained for Cronbach's  $\alpha$ . The Standard Error of Measurement values are all quite small implying an acceptable level of confidence in scoring precision. The failure rates are quite high, especially so for the Physics assessment.



Figure 1 shows an integrated frequency distribution<sup>2</sup> for the scores, with a theoretical, integrated Normal distribution plot calculated from the mean and standard deviation of the actual scores. The data are for Anatomy Set 31, but are typical of the Part 1 papers. Vertical lines give the raw cut score and the raw cut score plus one standard error of measurement ( $SE_M$ )<sup>3</sup>. The plots are taken from the analysis of the Part 1 Anatomy Set 31 examination, but are fairly typical of all five Part 1 examinations.

Table 1: Part 1 FRCR Clinical Radiology									
Part 1 Paper	No of candidates	No of items	Cut score, %	Mean score, %	Standard deviation, %	Skew	Cronbach's $\alpha$	Standard Error of Measurement ( $SE_M$ ), %	Percent failing at P/F cut score
Anatomy Set 25	58	100	73.00	78.68	11.09	-2.14	.903	3.45	22.41%
Anatomy Set 29	166	100	80.00	82.09	10.43	-1.46	.899	3.31	31.93%
Anatomy Set 30	77	100	76.00	81.15	9.75	-1.23	.889	3.25	23.38%
Anatomy Set 31	77	100	79.00	83.75	6.70	-0.71	.787	3.09	23.38%
Physics	482	199	74.87	77.15	9.39	-0.70	.919	2.67	35.48%



<sup>2</sup> The integrated frequency distribution is, as the name implies, the integral of the frequency distribution of the candidates' scores. It has the advantage of smoothing the frequency distribution plus normalising all cohort sizes to 100%. The point at which the 50% on the cohort axis intercepts the marks' curve gives, on the marks axis, the median score for the assessment.

<sup>3</sup> The raw cut score +1  $SE_M$  gives the 84% confidence interval for a true passing score.

It can be seen that the calculated Normal distribution is not a close fit to the actual scores distribution. The explanation is seen in the values of skew given in Table 1. These are all negative (in two cases substantially so) and this means that there is some clustering of scores at high values and a tail of scores at low values. Negative values for skew are common to many assessments and may be due to too many 'easy' questions or to a 'tail' of weakly performing candidates. Both effects produce divergence from the Normal distribution of scores as is seen here.

The Part 1 Anatomy assessment is run in four 'sets'. It looks as though each set is given a different assessment, although this point requires clarification. Since different candidates in the same cohort take a different set there could be an issue here if the difficulty of the paper varies for each set and this is not taken into account in setting the passing score. Ideally a standard setting procedure that takes account of test difficulty, or a test equating process should be used in such cases. Looking at the data in Table 1 for these four Anatomy exams it is clear that while they are similar in many respects, they are not identical. Analyses were therefore performed to see if the scores in the four sets differed from each other statistically. The analysis was performed in SPSS using ANOVA<sup>4</sup> with a Bonferroni test for significant differences. The results are shown in Table 2.

<b>Table 2: Comparison of the average scores on different Anatomy sets (* is significantly different at the 5% level).</b>			
<b>Set</b>	<b>25</b>	<b>29</b>	<b>30</b>
<b>29</b>	P = 0.13		
<b>30</b>	P = 0.88	P = 1.00	
<b>31</b>	P = 0.02*	P = 1.00	P = 0.59

Table 2 shows that the four Anatomy examinations do not generate significantly different mean scores except for the difference between Sets 25 and 31, where the mean scores are significantly different at the 2% level. This is a little worrying since ideally all the sets should be of equal difficulty and the mean scores similar. Of course this assumes that, given the substantial numbers of candidates, the average ability would be very similar on the four sets. The passing standards for the four sets also differ. This complicates the picture but despite Set 31 having a higher mean score the fail rates were broadly comparable across the four sets and none of the effects are large.

Although the four Anatomy sets differ somewhat in their means scores, etc., they can be merged into a single group and these scores can then be compared with the scores obtained by those candidates who also sat the Physics assessment. To look at concurrent validity (i.e. the relationships between two or more assessments taken at about the same stage of educational development) values for correlation coefficients were calculated. It was possible to do this because of the 482 candidates taking the Physics assessment; some 316 had also taken one of the Anatomy sets in the Spring 2014 set of examinations. Two methods were used, Pearson's *r*, which requires the two variables (Anatomy scores and Physics scores) to be Normally distributed, and Kendall's *tau*, a correlation coefficient for non-parametric data. This was done because, as Figure 1 shows, the scores for these assessments do not have perfect Normal distributions. The values obtained from the correlation between Anatomy

<sup>4</sup> SPSS is IBM's Statistical Package for the Social Sciences. ANOVA is an analysis of variance.

and Physics scores were  $r=0.62$  and  $\tau = 0.43$ . Both of these correlations are quite strong and both are statistically significant at the 5% level. This can be considered a good outcome since it shows that candidates who perform well in Anatomy also perform well in Physics and while the two tests assess different areas of knowledge and its application, some correlation between the areas is perhaps to be expected.

### 3.2.a Final Part 2A examinations held in Spring 2014

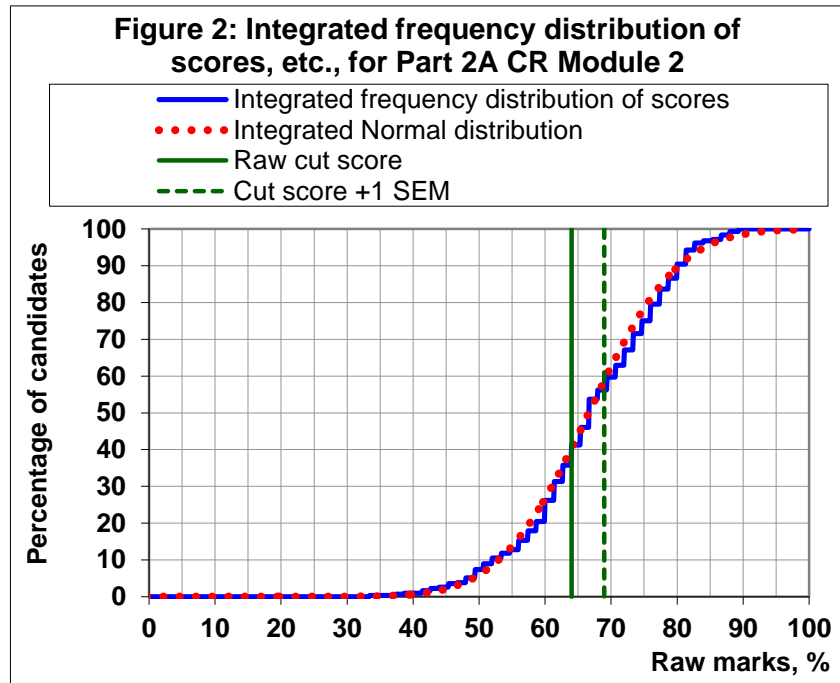
The Part A examinations in Clinical Radiology take the form of six, separate 75-item SBA papers covering the six modules of the Part 2 award. The basic whole-test-statistics for these examinations are given in Table 3.

Part 2A Paper	No of candidates	No of items	Cut score, %	Mean score, %	Standard deviation, %	Skew	Cronbach's $\alpha$	Standard Error of Measurement, %	Percent failing at P/F cut score
CR Module 1	343	75	62.67	58.44	10.43	-0.47	.763	5.08	65.60%
CR Module 2	397	75	64.00	62.74	9.60	-0.15	.714	5.13	52.64%
CR Module 3	313	75	64.00	66.98	10.64	-0.32	.782	4.97	35.78%
CR Module 4	323	75	64.00	64.11	9.48	-0.45	.717	5.04	42.41%
CR Module 5	307	75	62.67	63.08	9.71	-0.17	.723	5.11	50.49%
CR Module 6	348	73	61.33	63.42	11.01	-0.21	.782	5.14	43.10%

Taken as a group module examinations show modest mean scores (Range: 58.44% to 66.98%) with fairly large standard deviations (Range: 9.48% to 11.01%). All the papers produce negative skew values<sup>5</sup>, some of them quite large. Reliability, as judged from Cronbach's  $\alpha$  is around .72 to .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  $SE_M$  values are consistent at around 5%, but are a little on the large side. This is a consequence of the large standard deviations with low-ish values for  $\alpha$ . The passing standards are fixed and the failure rates seem very high. If Annex B is consulted it will show that four of the six module assessments have ten or more of the 75 items reducing the overall reliability of the test. The reasons for this should be explored with a view to improving test reliability.

<sup>5</sup> Negative values for skew imply clustering at higher scores and a 'tail' in the distribution at lower scores.

Figure 2 shows the integrated frequency distribution of the scores for Module 3 which is fairly typical of the six modules



The plots shows that the spread of marks is greater than in the Part 1 examinations and the Normal distribution plot follows quite closely the actual marks' distribution, even though these data have a small skew value of -0.32.

Although the six module assessments are different it is of interest to know the extent to which marks scored on the six tests differ statistically. Once again, to look at concurrent validity requires that scores for the same candidates are available for all of the six module assessments. Unfortunately, while up to 397 candidates took the module examinations, only 33 of these took all six during the Spring 2014 cycle of assessments. This sample is small for meaningful analysis and the outcomes should not be generalised to all candidates in the Spring 2014 pool. However, for this small group there are quite strong and statistically significant correlations between their scores on these 6 papers (Pearson's  $r$  range: 0.62 to 0.78; Kendal's tau range: 0.34 to 0.60, depending on which pairs of papers are compared). Cronbach's  $\alpha$  for overall scores on the six modules is high at .92, indicating a robust assessment profile for these 33 candidates. This is particularly good given that only 6 marks contribute to the assessment profile. It would be useful to make similar measures for all of the candidates taking Part 2A, although this would require merging marks for many different versions of the Part 2A module assessments. A large pool of candidates would be required to account for the different test contents and such an exercise would probably require a professionally constructed relational database containing individual item scores as well as total scores and candidates' demographics.

### 3.2b Final Part 2B examinations held in Spring 2014

The Part B examination in Clinical Radiology has three components; two comprise (1) 30 Rapid Reports, and (2) 6 Reports, both using images, etc.. Both also use double marked written responses. Raw marks are converted to marks in the range 4 to 8. The third component comprises two oral assessments using images with two pairs of examiners who generate marks on a scale from 4 to 8 points. Scoring is guided by anchor statements which provide descriptors for candidates' performance. The scores are combined into a single mark to which additional criteria are applied to determine the outcome. Where possible, the following analyses use raw and not adjusted marks, so the marks for rapid reports can in principle range from 0 to 30, those for Reports and Orals range from 4 to 8.. Table 4 shows basic whole test-statistics for these examinations.

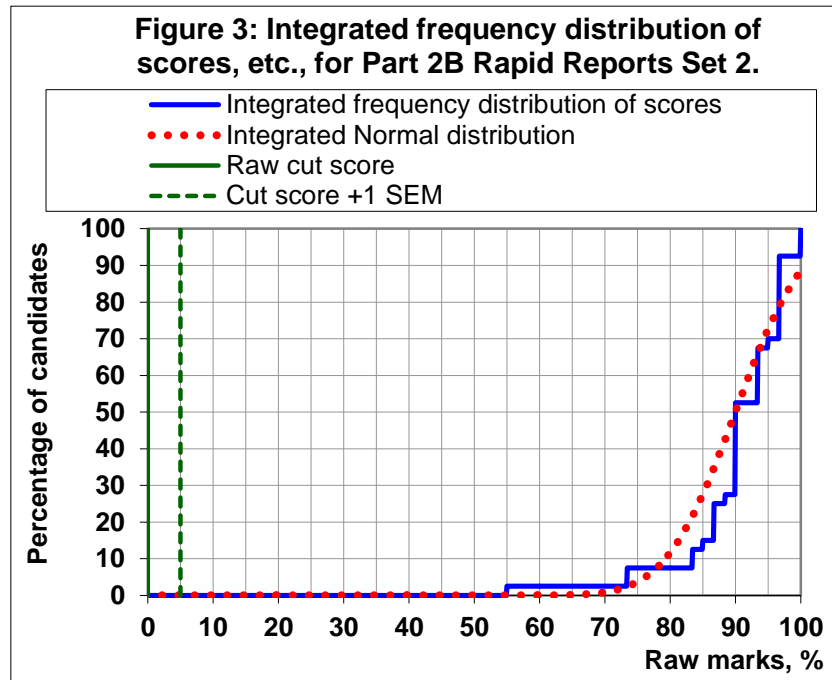
**Table 4: Part 2B FRCR Clinical Radiology**

Part 2B Assessment (Set)	No of candidates	No of items	Mean score, %	Standard deviation, %	Skew	Cronbach's $\alpha$	Standard Error of Measurement, %	Percent failing at P/F cut score
CR 2B Rapid Reports Set 1	76	30	91.07	7.33	-1.11	.564	4.84	n/a
CR 2B Rapid Reports Set 2	40	30	90.33	8.36	-2.29	.648	4.96	n/a
CR 2B Rapid Reports Set 3	78	30	88.63	6.21	-0.74	.266	5.32	n/a
CR 2B Rapid Reports Set 15	40	30	83.88	8.62	-1.19	.534	5.88	n/a
CR 2B Reports Set 2	39	6	74.13	5.16	0.07	.436	3.88	n/a
CR 2B Reports Set 4	40	6	75.52	5.69	-0.56	.541	3.86	n/a
CR 2B Reports Set 28	76	6	73.97	6.23	-0.12	.562	4.12	n/a
CR 2B Reports Set 29	78	6	71.44	6.42	0.07	.506	4.52	n/a
All four elements of Part 2B <sup>6</sup>	233	4	72.95	6.52	0.05	.565	4.30	54.94%

Both the Rapid Reports and the Reports are each run in four sets. Table 4 shows that the 30 item rapid reports' assessments have very high cohort mean scores, modest standard deviations as well as high degrees of skew towards high marks. More importantly they have and weak (to unacceptably low) reliability coefficients (Cronbach's  $\alpha$ ) for such high stakes assessments. The  $SE_M$ s are quite large in some cases indicating wide confidence intervals for the precision of the scores. The high means suggest that the assessments are perhaps too easy. In addition, more detailed test statistics in Annexes 1A and 1B show that for these papers there are quite low average item discriminations and that high numbers of items (in two tests, 16 out of 30) actually reduce overall test reliability (measured as Cronbach's  $\alpha$ ). Figure 3 shows the integrated frequency distribution of Rapid Reports Set 2, together with a

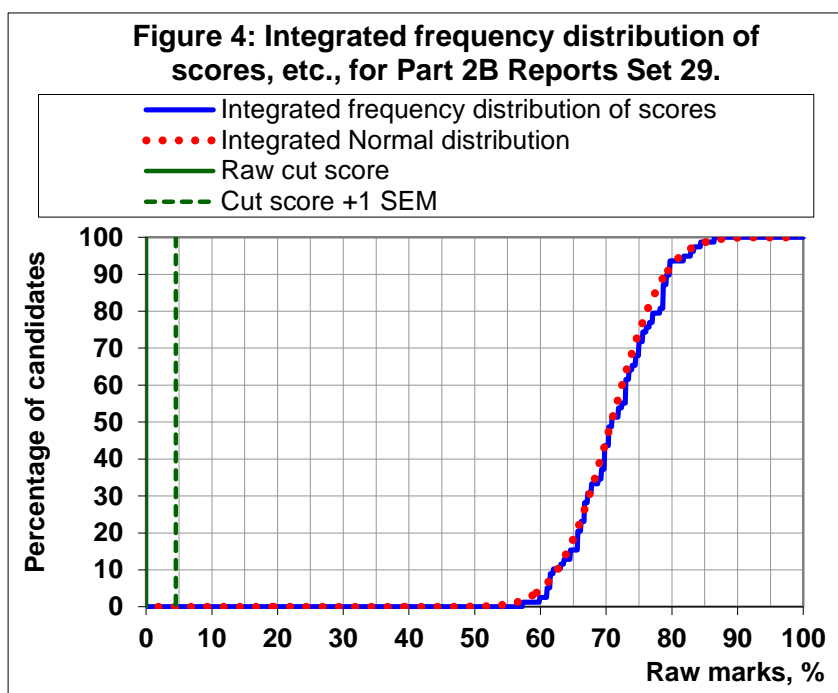
<sup>6</sup> These are the rapid reports, the reports and two oral examination scores treated here as separate scores (rather than the sum of the two orals). Raw scores are used in each case.

theoretical, integrated Normal distribution calculated from the mean and standard deviation of the actual scores.



The plot for the scores is shifted towards high scores and diverges markedly from the Normal distribution calculated for the mean and standard deviation of the actual scores. In fact the plot for the Normal distribution would not reach its maximum until a score greater than 100% had been achieved.

The four 'sets' of Reports, each set comprising six reporting tasks, have general statistics similar to the rapid reporting cases, although the means are more reasonable (lower), the standard deviations are smaller and the data show little or no negative skew in the scores' distributions. However, the values for Cronbach's alpha are again too low. This may be because there are only six 'samples' of performance in the Reports assessments. Cronbach's alpha does produce high values with small numbers of test items unless internal consistency of scores is very high. The small number of test items is the likely cause, since Annex 1B shows that these tests have reasonable item to total correlations and that no items reduce overall reliability. Average discrimination is, however, on the low side. Improved reliability might simply require a greater number of items. The SEMs are reasonable, but ideally could be smaller. Figure 4 shows the integrated frequency distribution of Reports Set 2, together with a theoretical, integrated Normal distribution calculated from the mean and standard deviation of the actual scores. The scores follow the Normal distribution quite closely, showing this test to have quite different properties to the Rapid Reports assessments. This difference may reflect the different mark schemes used in the two assessment formats.



While the two oral examinations each with a pair of examiners generate four scores, only the average score for each Oral is used here. With only a small sample of examiners' scores it is only useful to provide means and standard deviations of cohort scores. Conventional reliability analysis is not particularly useful, (except perhaps inter-rater correlation, which has not been done for this report) because with only two oral scores per candidate, the reliability as measured with Cronbach's  $\alpha$  will be very low.

The bottom row in Table 4 shows the statistics for the four elements of Part 2B combined. The four elements are: the Rapid Reports and the Reports plus the marks from the Oral 1 and Oral 2 assessments. For each Oral the average of the two examiners' scores has been used. Adjusted (close) marks have been used for all four elements rather than the raw marks for some other analyses and the two Oral scores kept separate and not merged. The average score for the cohort at 72.95% is reasonable; the standard deviation is quite small at 6.52%. There is very little skewing of scores to the high end of the range. Cronbach's  $\alpha$  at .57 remains sub-optimal but this may in part be due to only four scores contributing to the reliability calculation, which will reduce  $\alpha$ .

As for the Part 1 Anatomy assessments, Part 2B is run as a series of sets, with presumably different test items in each set. It is therefore of interest to see if the four groups of candidates produced comparable scores since the passing standard is the same for all sets in Rapid reports and Reports assessments. To do this an ANOVA analysis with a Bonferroni test was performed in SPSS to look for significant differences between the mean scores of these tests. The results are shown in Tables 5 and 6. These show that for the rapid reports, Set 15 had a statistically significantly lower mean score than the other three sets. For the Reports, Set 29 had a statistically significantly lower mean score than both Set 4. Of course the differences in mean scores may be due to differences in cohort ability, but

as the difficulty of each Set is not known, either by standard setting or test equating techniques, this may not be the case.

<b>Table 5: Significance of difference between mean scores on Rapid Reports sets (*significantly different).</b>			
<b>Set</b>	<b>1</b>	<b>2</b>	<b>3</b>
<b>2</b>	P=1.00		
<b>3</b>	P=0.254	P=1.00	
<b>15</b>	P=0.00001*	P=0.0008*	P=0.0068*

<b>Table 6: Significance of difference between mean scores on Reports sets (*significantly different).</b>			
<b>Set</b>	<b>2</b>	<b>4</b>	<b>28</b>
<b>4</b>	P=1.00		
<b>28</b>	P=1.00	P=1.00	
<b>29</b>	P=0.14	P=0.004*	P=0.06

In terms of concurrent validity, the correlations between Rapid Report scores and Report scores when raw scores for the four sets are merged was weak but statistically significant when measured using Pearson's r ( $r=0.20$ ,  $P=.002$ ) and Kendall's tau ( $0.14$ ,  $P= 0.002$ ). The correlation seems rather weak if these two assessments are testing similar material.

A similar study was performed for all four elements (Rapid Reports, Reports, Oral 1 and Oral 2) but using the adjusted (close) marks. This produced the following correlation matrix (Table 7) using Kendall's tau measurements. As might be expected, the highest correlation is between the two oral examinations, although the correlation is still quite weak. All the correlations are significant at the 1% level, but in general the correlations are modest in size. Slightly higher values for correlation are obtained if Normal distributions are assumed and Pearson's r is used to test for correlation ( $r$  range: 0.168 to 0.402).

<b>Table 7: Correlation matrix for elements of Part 2B. using Kendall's tau</b> (* = statistically significant correlation).				
	<b>Rapid reports</b>	<b>Reports</b>	<b>Oral 1</b>	<b>Oral 2</b>
<b>Rapid reports</b>	1.00			
<b>Reports</b>	0.133*	1.00		
<b>Oral 1</b>	0.227*	0.260*	1.00	
<b>Oral 2</b>	0.145*	0.236*	0.311*	1.00



## Fellowship of the Royal College of Radiologists in Clinical Oncology

### 4.1 Part 1 Examinations held in Spring 2014.

Table 8 shows the basic, whole-test statistics for these examinations. Three of the four papers in Part 1 CO have similar characteristics with reasonable cohort mean scores in the range 64.5% to 66.4% while the fourth paper, Physics, has a rather lower mean at 57.4%. 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 for Cronbach's  $\alpha$  (range .831 to .911), although this may be due to the large standard deviations, which can inflate values for  $\alpha$ . Similarly the large standard deviations produce quite large values for the Standard Error of Measurement despite the high values for  $\alpha$ . The failure rate is high for the Cancer Biology paper and, at 49%, very high for the Physics paper. All papers show some degree of negative skew in the marks' distribution, but this is only large for Medical Statistics. The papers comprise 40 to 50 MCQ questions, and item analyses in Annex 2 show that in most cases around 6 items detract from overall test reliability. However, this is not a large number and overall reliability as well as other metrics are good.

**Table 8: Part 1 FRCR Clinical Oncology**

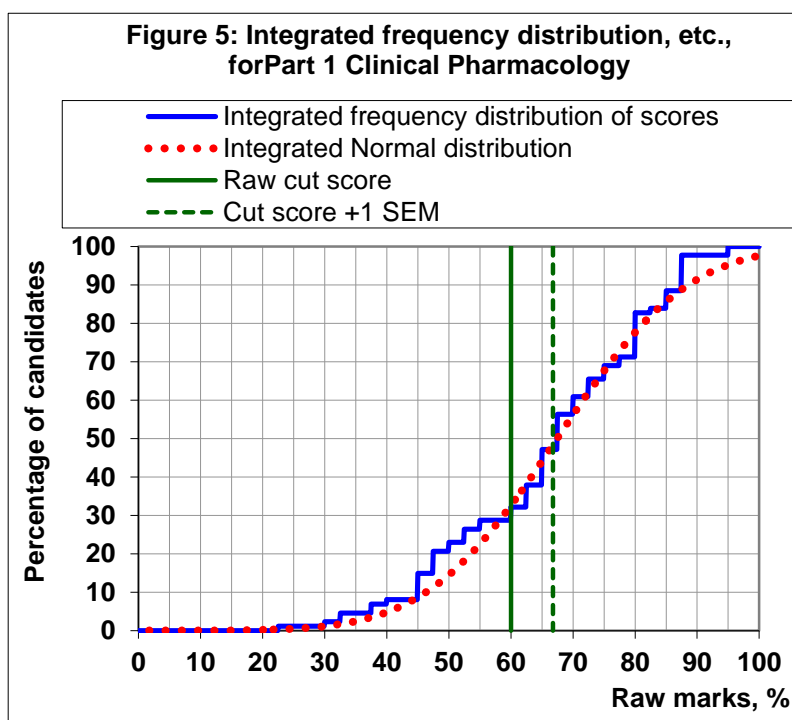
Part 1 FRCR Paper	No of candidates	No of items	Mean score, %	Standard deviation, %	Skew	Cronbach's $\alpha$	Standard Error of Measurement ( $SE_M$ ), %	Percent failing at P/F cut score
Cancer Biology / Radiobiology	60	50	64.47	19.88	-0.31	.911	5.93	38.33%
Clinical Pharmacology	87	40	65.78	16.49	-0.44	.831	6.78	28.74%
Medical Statistics	87	40	66.41	18.98	-0.89	.875	6.71	24.14%
Physics	53	50	57.43	19.00	-0.26	.894	6.18	49.06%

The graphical plots in Annex 2A for these papers show that the scores diverge to varying degrees from a calculated Normal distribution, but in other respects Figure 5, for the Clinical Pharmacology scores, is fairly typical of the group. Figure 5 shows an integrated frequency distribution<sup>7</sup> for the scores, with a theoretical, integrated Normal distribution plot calculated from the mean and standard deviation of the actual scores. Vertical lines give the raw cut score and the raw cut score plus one standard error of measurement ( $SE_M$ )<sup>8</sup>

<sup>7</sup> The integrated frequency distribution is, as the name implies, the integral of the frequency distribution of the candidates' scores. It has the advantage of smoothing the frequency distribution plus normalising all cohort sizes to 100%. The point at which the 50% on the cohort axis intercepts the marks' curve, gives on the marks axis, the median score for the assessment.

<sup>8</sup> The raw cut score +1  $SE_M$  gives the 84% confidence interval for a true passing score.

It can be seen that the calculated Normal distribution is not a close fit to the actual scores' distribution. The differences between the two plots in this and other Part 1 papers are mainly due to the degree of distribution skew, for which values are given in Table 8. The skews are all negative (but only large for Medical Statistics). This means that there is some clustering of scores at high values and a tail of scores at low values. Negative values for skew are common to many assessments and may be due to too many 'easy' questions or to a 'tail' of weakly performing candidates. Both effects produce divergence from a Normal distribution of scores as is seen here. Detailed inspection of item-level statistics may help reveal the cause if due to individual item performance. These metrics are not given here but are available in Annexes 2A and 2B.



The numbers of candidates sitting the various Part 1 papers in Spring 2014 ranged from 53 to 87. Of these, 37 candidates took all four papers at the same time. It is therefore possible to examine concurrent validity for the four Part 1 examinations in Clinical Oncology. Because the scores diverge from Normal distributions, a non-parametric test of correlation (Kendall's tau) was performed on these candidates' scores using SPSS. This gave the correlation matrix shown in Table 9. This shows quite strong correlation between the scores on the four papers which suggest that in validity terms the papers are sampling different but closely related aspects of the same subject area. This is good as is the fact that all the correlations are statistically significant at the 1% level.

<b>Table 9: Correlation matrix for elements of Part 1. using Kendall's tau (** indicates statistically significant at the 1% level)</b>				
	<b>Cell Biol. / Radiobiol.</b>	<b>Clin. Pharm.</b>	<b>Med. Stats.</b>	<b>Physics</b>
<b>Cell Biol. / Radiobiol.</b>	1.000			
<b>Clin. Pharm.</b>	0.634**	1.000		
<b>Med. Stats.</b>	0.614**	0.540**	1.000	
<b>Physics</b>	0.674**	0.583**	0.690**	1.000

The assessment profile of four papers in Part 1 has a Cronbach's  $\alpha$  derived from the final percentage scores of .93, a very high value given that there are only 4 marks for each candidate in this particular analysis.

#### 4.2.a Final Part 2A examinations held in Spring 2014

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.

Paper	No of candidates	No of items	Mean score, %	Standard deviation, %	Skew	Cronbach's $\alpha$	Standard Error of Measurement, %	Percent failing at P/F cut score
Part 2A	52	239	60.66	7.45	-0.23	.856	2.83	48.08%

Table 10 presents the basic summary statistics for this examination. It can be seen that the mean cohort score is modest at 60.7%, the standard deviation is a quite small 7.4%. The examination has good reliability, with a Cronbach's  $\alpha$  of .86 and a small standard error of measurement at 2.83%. The examination has therefore performed well, although at 240 items, perhaps even higher reliability might occur. If the detailed analysis in Annex 2 is consulted it will be seen that 35 of the 239 items actually lowered overall reliability. While in large-scale SBA examinations some items will often show this property, the number here does seem high. One item was withdrawn from the total following analysis of the results and the resulting pass mark (a fixed standard, adjusted for deletion of items from the test) was 60.42%. Since the mean score was 60.7, it is not surprising that 48% of the cohort failed this assessment. The integrated frequency distribution of the scores is shown with other information in Figure 6.

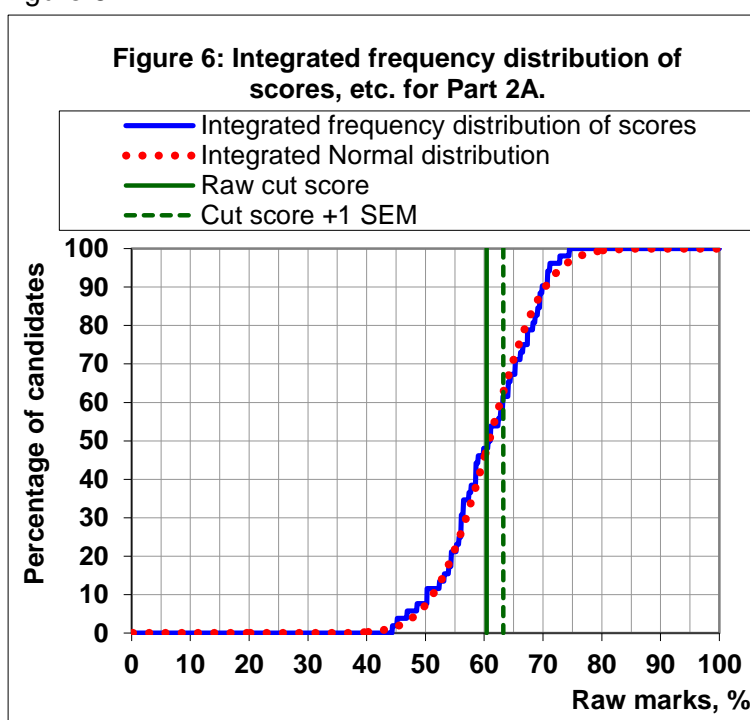


Figure 6 shows good agreement between the plot for candidates' scores and that for a Normal distribution calculated from the mean and standard deviation of the score. The marks' distribution plot fits well the Normal distribution plot, probably because of the low skew value of -0.23.

#### 4.2.b Final Part 2B examinations held in Spring 2014

There are two assessment components for Part 2B in Clinical Oncology. The first is a 5-station clinical examination using real patients, the second an 8-case, on-screen structured oral examination using images and reports, etc.

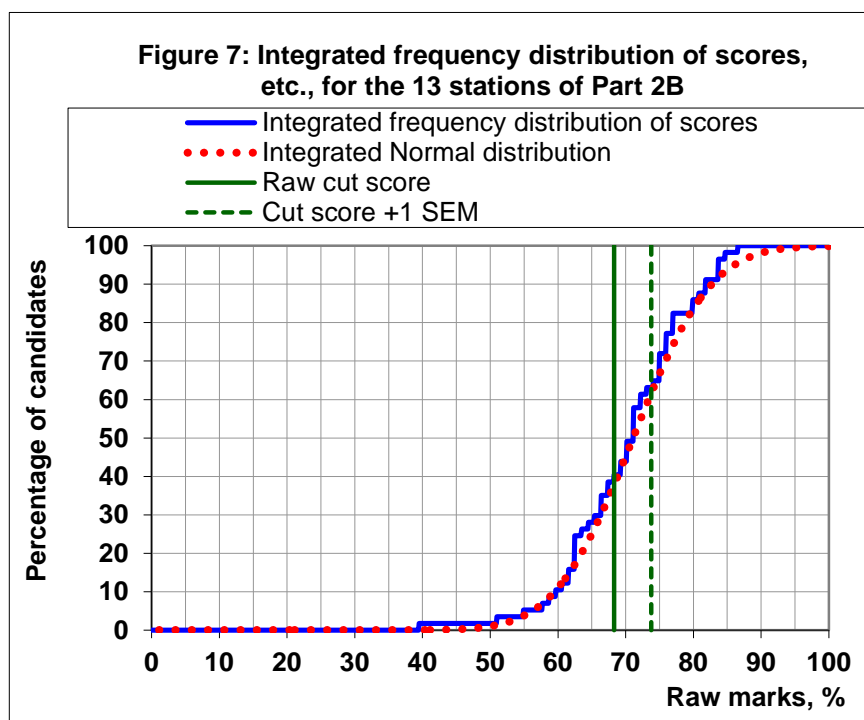
Each examiner scores each candidate independently on a four-point scale (performance anchor statements are provided) and pairs of examiners' scores are summed. The total score for the 13 stations, together with other criteria (e.g. the number of stations failed) are used for pass-fail decisions. The pass mark for each station is fixed at 6 out of 8 marks. See also footnote<sup>7</sup>. A summary of the basic statistics for the two components and the overall scores for Part 2B is given in Table 11.

Table 11: Part 2B FRCR Clinical Oncology								
Part 2B component	No of candidates	No of items	Mean score, %	Standard deviation, %	Skew	Cronbach's $\alpha$	Standard Error of Measurement, %	Percent failing at P/F cut score <sup>9</sup>
Clinical (5 Stations)	57	5	70.18	10.85	0.12	.289	9.15	n/a
Structured Oral (8 Stations)	57	8	69.98	11.12	-0.54	.632	6.75	n/a
CO 2B Overall (13 stations)	57	13	70.06	9.07	-0.63	.631	5.51	40.35%

The mean scores are quite high at around 70% and the standard deviations are quite large (9.0% to 11.1%). The values for the skews of the marks' distributions are variable showing some clustering at high scores in the 8-station clinical examination and therefore in the overall marks. The overall reliability is .63 as measured by Cronbach's  $\alpha$  and is also .63 for the 8-station clinical component. These are low values, but it can be difficult to obtain higher values for clinical assessments. Ideally a value of .75 would be more acceptable. The value for Cronbach's  $\alpha$  for the clinical component of .29 is unacceptably low, even when allowance is made for a sample of only five items contributing to the calculation. Standard errors of measurement are quite large (very large for the Clinical examination) and the failure rate is high at 40.4%.

<sup>9</sup> Please note that the overall fail rate cited here is based on candidates achieving 71/104 marks. Other pass-fail criteria have not been included in this calculation.

Figure 7 shows the integrated frequency distribution and other plots for the full, 13-station Part 2B examination. There is some divergence of the scores' plot from the Normal distribution plot. The negative skew value of -0.63 indicates some clustering at high scores and a small tail at lower scores. If the plots for raw cut score and cut score plus 1 SEM in Figure 7 are compared with those for Part 2A (Figure 6) it can be seen that the higher SEM for 2B means that less confidence can be placed in scores around the pass-fail boundary. To have 84% confidence that all passes were 'true' passes (at the +1 SEM value) would mean that 63% of candidates would fail this assessment.



Part 2B assessments are not based on a single 'set' of identical questions (as assumed here for Parts 1 and 2A). It has not been possible to compare how different versions of Part 2B compare within the same diet of examinations. Concurrent validity can be examined by analysing the correlation between the Clinical and Structured Oral stations in FRCR(CO) Part 2B. This has been done using Pearson's  $r$  and Kendall's tau correlation coefficients in SPSS. Correlation between clinical and oral components' scores is weakly positive (Pearson's  $r = 0.314$  and Kendall's tau = 0.202; both are significant at the 5% level). Figure 8 shows a scatter plot of the scores. This suggests that Oral and Clinical components examine similar, but not identical, abilities.

Because only 16 of the more than 50 Part 2 candidates sat Part 2A and 2B in Spring 2014, it is not possible to examine correlation between Part 2A and 2B scores. Nor is it possible to compare performance at Part 1 and 2 to evaluate predictive validity for the candidates.

