

JCHMT

JOINT · COMMITTEE · ON · HIGHER · MEDICAL · TRAINING

Knowledge-Based Assessment Pilot Project

Final Project Report

October 2006

Contents

1	Introduction	3
2	Principles.....	3
3	Organisation	5
4	Question / Exam Setting Process	6
5	Pilot Operation	7
6	Results	11
7	Question Bank.....	22
8	Discussion.....	23
9	Conclusions.....	26
10	Acknowledgements	27
11	Appendix A - Cardiology Details	28
12	Appendix B - Dermatology Details	29
13	Appendix C - Gastroenterology Details.....	30
14	Appendix D - Geriatric Medicine Details	31
15	Appendix E - Neurology Details	32

1 Introduction

In recent years there has been much work within JCHMT, the Colleges and medical education generally, to investigate and implement new methods of assessment. A successful pilot study of three methods of performance assessment for Specialist Registrars (SpRs) was completed by the Royal College of Physicians in 2004 and these methods are now being rolled out. A JCHMT project was started in 2004 to pilot the use of knowledge-based assessment for SpRs and was completed in 2006 with the writing of this report.

The philosophy behind this project lay in the increasing expectations of patients, the public and Government, that specialists dealing with them have the appropriate knowledge of the specialty in which they propose to practise as independent practitioners. Additionally, one of the three dimensions built into the subject matter of the competence-based curricula, knowledge, is only partially addressed by the three pilot assessment methods.

The objectives of the pilot were:

- To consider the reliability and validity of the proposed assessment method in this context.
- To understand the practical implications of introducing knowledge-based assessment

Four specialties initially agreed to participate: Cardiology, Gastroenterology, Geriatric Medicine and Neurology. A fifth, Dermatology, joined the project later. Gastroenterology withdrew before the end of the project and did not set a paper – reasons for this are discussed in section 8.1 and Appendix C.

2 Principles

Five project principles were established before the project started, though these were to be modified as the project progressed:

- Assessment to be formative with a summative element
- Will use “best of five” format, multiple choice questions
- Will be web-based
- Will be centred on the core elements of specialty curricula.
- Will generally be taken in year 2 (or different years where modular curricula used).

2.1 *Formative / Summative*

The original vision was of a situation where trainees would have access to an online bank of questions which could be used formatively, at any time or place, to guide their learning, but that at some stage they would be required to take a summative test in a secure, invigilated setting. During early discussions with the specialties it quickly became apparent that their interest was mainly in summative assessment and also that there would be a practical problem in creating enough questions to use for both formative and summative assessment. It was agreed with each specialty that the assessments should be “closed-book”, i.e. with no access to reference materials, and taken under examination conditions.

2.2 *Best-of-Five*

Best-of-five format questions have replaced true/false questions in MRCP(UK) parts 1 and 2 in recent years, so were an obvious choice for this project. The Colleges have built up considerable expertise in writing this type of question.

Best-of-five questions consist of a descriptive stem, a specific lead-in question and five options from which examinees must choose one best answer – see the example below:

An 82-year-old woman presented with a right foot drop of several months' duration. Examination revealed markedly weak right ankle dorsiflexion (MRC grade 2/5), normal plantar flexion and eversion, mildly weak inversion, and absent extension of the hallux. Sensory examination was normal.

Investigations:

MRI lumbar spine	normal
EMG / NCS	denervation in L4/5 territory muscles normal superficial peroneal sensory action potential.

At follow up 6 months later the weakness had progressed and in addition there was mild weakness of plantar flexion.

What is the most likely diagnosis?

- A common peroneal nerve palsy
 - B L5 radiculopathy
 - C malignant lumbosacral plexopathy
 - D mononeuritis multiplex
 - E motor neurone disease
- (E is the correct answer)

Unlike true/false questions, options do not have to be unequivocally true or false but one option must clearly be the best answer, so in the example above any number of the options might be possible diagnoses but one must be much more likely than the other four.

There was some consideration given during the project to allowing alternative question types, such as “n from many” and extended-matching questions, but it was agreed that these did not offer any significant advantages to offset the complexity of managing different question types.

2.3 Web-Based

Once the idea of constantly accessible formative assessment had been abandoned there was a much less compelling case for using web-based assessment. Eventually practical decisions on how best to run the pilot assessments meant we reverted to a paper-based examination.

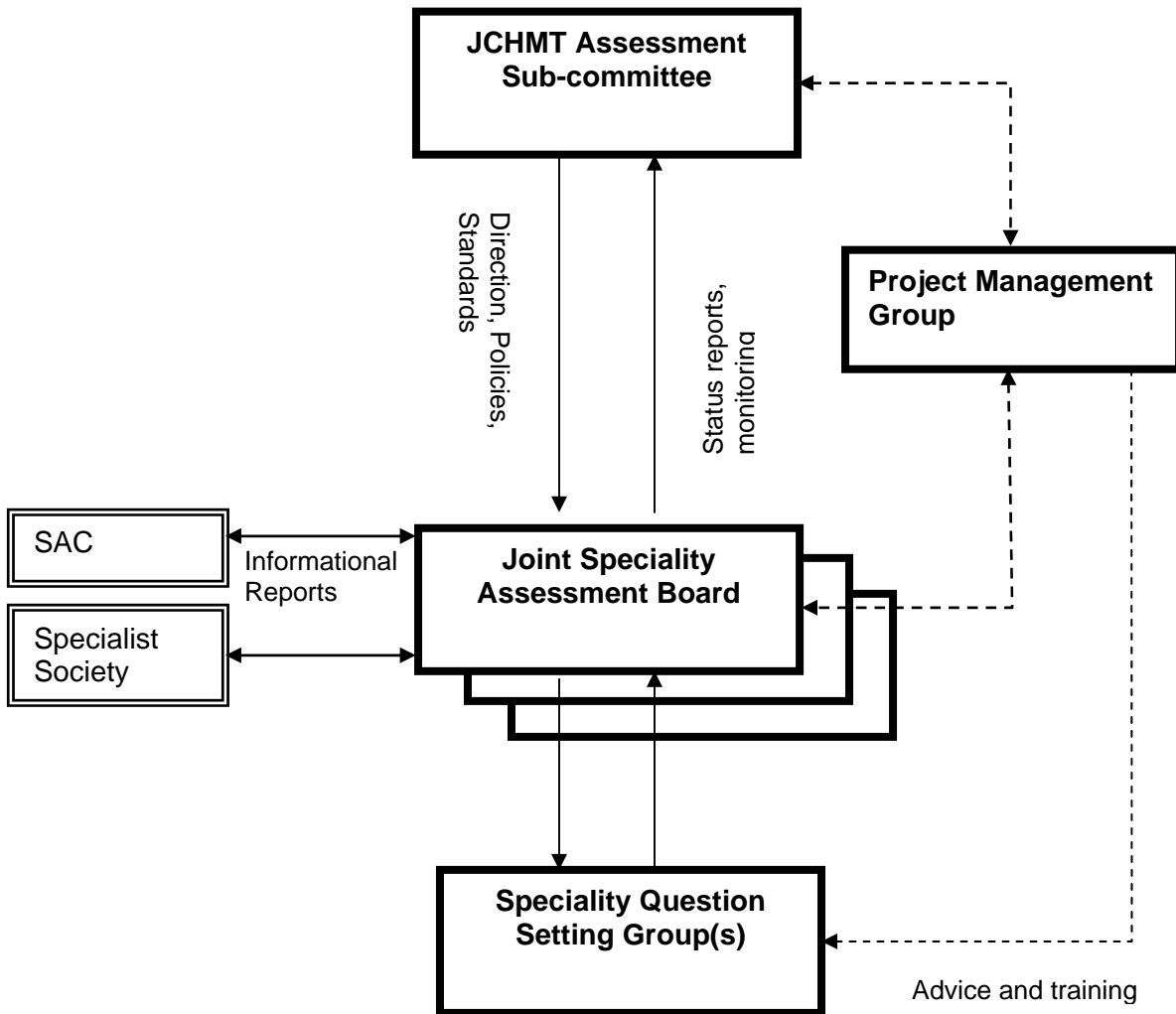
2.4 Syllabus

Question writers were asked to restrict themselves to core components of the Higher Medical Training curriculum for the specialty. Participants were told that questions would be based on these curricula. Some specialties discussed specifying reading materials or resources for participants, but decided against this.

2.5 Year of Training

Specialties in the pilot generally anticipated that, if implemented beyond the pilot, they would expect SpRs to take assessments from around the end of year 2 or 3, with the option of retaking as necessary in order to pass before penultimate year assessment. In practice there would probably be no restriction on someone choosing to take it earlier, following the start of specialist training.

3 Organisation



3.1 JCHMT Assessment Sub-Committee

This committee met 3 times a year and was responsible for steering and monitoring performance and knowledge-based assessment projects. It provided oversight and guidance to the project.

3.2 Project Management Group

This group met as required throughout the project (monthly at some times, less frequently at others) to monitor progress and take practical decisions on direction. Representatives from the specialties attended some meetings. The project manager was the link from this group to the specialties.

The group comprised:

JCHMT Medical Director
 JCHMT General Manager
 JCHMT Deputy General Manager
 RCP Director of Education
 Project Manager

Dr George Cowan / Dr Chris Clough
 Mr Nicholas Grant
 Ms Lesley Hagger
 Mrs Winnie Wade
 Mr Joe Booth

3.3 Joint Specialty Assessment Board

A project board was established for each speciality, with responsibility for knowledge-based assessment within that speciality. In some cases this was an extension of a pre-existing committee (e.g. the Neurology Training and Assessment Committee), but all included representation from both the SAC and the relevant specialist society. The committee chairs were the main point of contact for the project manager.

4 Question / Exam Setting Process

The process that was followed for setting questions and the pilot exam papers was generally modelled on that used for MRCP(UK). There were some differences by speciality which are expanded on in the appendices to this report.

4.1 Question Writing

Writing good questions to a consistent style and level is not easy

Standards for the type, style, format, wording etc. of questions were created based on the standards in place for MRCP(UK). These were supported by templates, examples and tips. This standards document was given to all question writers. During the course of the project a new MRCP(UK) question writing manual became available. This excellent, comprehensive document was given to some question writers who attended meetings and used as a reference work.

Question writers were identified by a combination of personal approaches from board members and more general requests for volunteers e.g. through society mailings and meetings. It was very helpful to have a number of writers with experience of writing for MRCP(UK). Most question writers were consultants but a few SpRs were involved in the process, either in their own right or contributing via consultant colleagues.

Writers were invited to Question Setting Days at the College. Standards and templates were sent out and writers were asked to contribute 10 questions each in advance.

The Question Setting Days typically started with brief presentations from practitioners on how to write good questions, then split into small groups to review and edit each other's questions. Questions were either accepted, rejected completely, recommended to be reworked, or edited during the meeting into acceptable form. Very few questions were accepted without some editing. This was generally effective as a good way of training people on question-writing whilst also generating usable questions.

Accepted questions were categorised and stored in a database in a common form.

Attendance at question setting days was approved for external CPD points, in the same way as MRCP(UK) question setting, up to a total of 12 external credits in one year. Further time could be counted as personal credits. Question setting work not linked to attendance at a meeting can be claimed as personal CPD.

4.2 Paper-Setting

One person made a draft selection of questions to be included in the paper, using a blueprint of topics to ensure appropriate weighting by subject area. This draft paper was reviewed by a group of board members and question setters. During this review some questions were rejected as being either inappropriate for the paper or unusable in future and replacement questions were selected. Other questions were edited again at this stage.

The final, formatted paper was proof-read by a further reviewer who, in most cases, had not been on the selection group.

4.3 Standard Setting

Three specialties decided to set a notional pass mark, based on the Anghoff method, in advance of candidates taking the assessment. This gave an indication in advance of how well participants were expected to perform and allowed for comparison of actual performance against this.

A standard setting group of about 6 people was established. The selected questions for the pilot were sent to each standard setter in advance. Each standard setter rated each question according to what percentage of just-passing candidates should get this question right. Ratings were collated into a spreadsheet, and means, ranges and standard deviations calculated for each question.

A standard setting meeting was held at which each question was briefly discussed and raters asked to support their views, particularly those at the extremes of the range. After discussion, each group member was then asked to re-rate the question. These ratings were recorded, the mean calculated for each question, and the mean of these taken to give the pass-mark.

One of the problems with this approach was establishing a common idea of what a “just-passing candidate” is like, i.e. someone who is at the borderline of being just good enough to pass. One has to have a concept of what level of “just passing candidate” one is talking about. In neurology this was explicitly agreed to be the newly achieved CCST trainee. For dermatology it was based on the core dermatology that it was judged that an SpR should know at the start of year 3 of training.

5 Pilot Operation

In order to maximise participation and minimise cost and disruption to participants it was agreed to hold sittings of the pilot assessments locally in all regions (deaneries) of the UK. For each specialty a local organiser was found. Approaches were made to Regional Specialty Advisors, Programme Directors or chairs of Specialty Training Committees. In many cases these people took on the lead role themselves but some delegated the job to colleagues. In a few regions deanery administrators took a major part in organising facilities and contacting trainees. In most cases there was one location per region, but some specialties/regions arranged for multiple locations to be used. For 3 specialties a room was booked at the Royal College of Physicians and made available as a central venue for London/Thames region participants.

The pilot assessments took place on different days for the different specialties, but as much as possible took place simultaneously within each specialty.

5.1 Venues

Organisers were asked to find a room or rooms, with tables and chairs, to hold the maximum number of participants. All other materials were supplied. The instruction was that ideally tables should be laid out in “exam format” all facing the front and with a gap between them. Although there was little incentive to cheat in the pilot the aim was to make it realistic.

The project refunded any costs incurred in booking rooms (though this was quite rare). No travel expenses were paid to participants.

5.2 *Invigilators*

Invigilators were asked to prevent any cheating, copying or collaboration and to ensure that question papers were not copied or taken away. Organisers were asked to ensure that at least 1 invigilator was a senior doctor (preferably the Programme Director, RSA or other senior consultant) to ensure proceedings were taken seriously

5.3 *Participants*

Organisers were asked to ensure that their trainees and consultants knew about the date and location and to encourage them to take part. All SpRs (NTNs and LATs) were strongly encouraged to take part, along with consultant volunteers, to provide a range of data. The neurology pilot was restricted to consultants within 5 years of award of CCST in order to check the validity of the standard setting process by using people who had not long successfully completed training. Other specialties were less restrictive and a broader range of consultants took part.

All trainees in the specialties were written to, via e-mail or post, using details from the JCHMT database, but the primary route for detailed communication was through local organisers. Information was put on the JCHMT web-site. Word was also spread through specialist society web-sites, newsletters and meetings.

Sample questions were made available (via organisers and the web-site) to give participants an idea of the nature of questions.

We guaranteed anonymity to participants by allocating random identifying numbers to them. Examination numbers were printed onto cards. These were shuffled and distributed at random to the regional centres (i.e. numbers were not sequential or grouped within a region). Organisers were asked to allocate the cards at random to participants and not to keep a record of these. In the sittings held at the RCP this was done by putting a card on each desk in advance and allowing participants to choose their seat. Candidates were subsequently to identify their result using their unique examination number.

5.4 *Printed Materials*

The exam papers were printed in-house as double-sided A4 documents bound within coloured card covers, with instructions to participants printed on the front. The font used was Arial 12-point. A large-print paper was produced for one participant.

Answer sheets were acquired via the MRCP(UK) Central Office. These were optically-readable, double-sided sheets allowing participants to select options A to E for 100 questions, identical in structure to the sheets used for MRCP(UK) P2 examinations (but without an MRCP heading).

Questionnaire forms were produced (by the RCP Medical Workforce Unit) to accompany the answer sheets in order to gather demographic data about participants, notably their role (SpR-NTN, SpR-LAT, consultant, other); year of training or years as a consultant; and any comments on this type of assessment or the specific paper.

5.5 *Distribution*

Packs were prepared for each regional centre containing everything needed for the pilot for the number of participants that organisers expected. This was: instructions to organisers; question papers; answer sheets; questionnaire forms; 2B pencils & erasers; pens; examination number cards and a return address label.

Packs were sent out to organisers a week in advance via overnight courier delivery. Organisers were asked to return all used and unused question papers, answer sheets and questionnaires. Couriers were initially booked to collect packs from the same locations the day after the pilot had taken place. For the first two pilots there were a number of problems with couriers not being able to find the collection points, or not being allowed to get to them. For the last two pilots packs were sent out by Royal Mail special delivery and organisers were given the option of returning them the same way or using a pre-booked courier.

Returned question papers were sent for secure shredding.

5.6 Marking

Answer sheets were scanned and marked automatically using MRCP(UK) equipment. Each question was worth a single mark for the correct answer, with no marks given for any alternatives and no negative marking. There were a few cases where 2 answers had been selected for a question and it was not possible to make a judgement on the participant's intention – these were scored as 0.

Questionnaire forms were scanned by optical character reading software and manually corrected where handwriting was flagged as illegible or ambiguous. Most of the participants' free-text comments had to be manually transcribed.

All SpRs were allocated to a year of training where this could be determined from the demographic data completed. All LATs were included in year 1, regardless of what they had recorded for year of training (e.g. some people recorded themselves as year 3 or 4 LATs).

Participants who categorised themselves as being in "other" roles (i.e. not SpRs or consultants) or who could not be allocated to a year of training were excluded from the comparative analysis.

5.7 Analysis

Results were analysed using a combination of Excel spreadsheets and SPSS statistics software, with invaluable support from the MRCP(UK) Central Office.

Participants were allocated to deciles according to their performance within their year/post group and also within the total population of participants (e.g. decile 1 being the bottom 10% of the cohort, decile 10 the top 10%).

Statistics were produced for every question showing the mean score (i.e. what proportion of people got it right); the correlation between participants' performance on the question and their performance on all the other questions in the paper (the point bi-serial, a number between -1 and 1); a breakdown of responses to each question by quintiles of candidates. This information is being used to inform decisions about which questions might be reused in future and to give guidance on what types of questions worked well or badly.

The example question shown in section 2.2 above was used in the neurology paper and the table below shows the actual performance of this question:

	quintiles					
	1	2	3	4	5	Total
blank	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
A	10.0%	2.2%	2.4%	1.9%	0.0%	3.1%
B	20.0%	15.6%	9.8%	7.4%	0.0%	10.2%
C	22.5%	31.1%	19.5%	13.0%	8.7%	18.6%
D	15.0%	0.0%	0.0%	5.6%	2.2%	4.4%
E	32.5%	51.1%	68.3%	72.2%	89.1%	63.7%
	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

To analyse the performance of the questions, participants were divided into 5 roughly equal groups (quintiles) according to their scores on the exam. Quintile 1 represents the 20% of participants who scored lowest on the whole paper, while quintile 5 is the top 20% of participants. Only 32.5% of quintile 1 selected the right answer (E) for this question, with the other 4 options all being selected by a number of participants. This gradually improved through the quintiles with 89.1% of quintile 5 getting it right, and only option C and D garnering a few votes. This upward slope of performance through the quintiles is indicative of a question that is working well to discriminate between candidates. The point bi-serial for this question was 0.367.

5.8 Feedback

Results for each specialty were published on the JCHMT web site, about 1 month after each pilot had taken place: www.jchmt.org.uk/assessment/knowledgeAssessment.asp

There was some variation by specialty in exactly what was published, but in each case there was some commentary and summary information and a full list of individual scores ordered by exam number.

Trainees and organisers were notified by e-mail as soon as the results were published.

6 Results

6.1 Participation

6.1.1 Number of Participants

	SpR1	SpR2	SpR3	SpR4	SpR5	Cons	Other	Total
Cardiology	61	83	60	51	25	12	11	303
Dermatology	46	53	53	29	n/a	33	6	220
Geriatrics	83	79	90	78	72	37	9	448
Neurology	47	33	43	39	34	30	0	226

6.1.2 SpRs registered with JCHMT (July 2006)

	ATN/FTN/FTTA	LAT	NTN	VTN	Total
Cardiology	1	41	452	48	542
Dermatology	3	17	171	26	217
Geriatrics	1	13	379	131	524
Neurology	2	18	160	22	202

6.1.3 Participation rates for SpRs

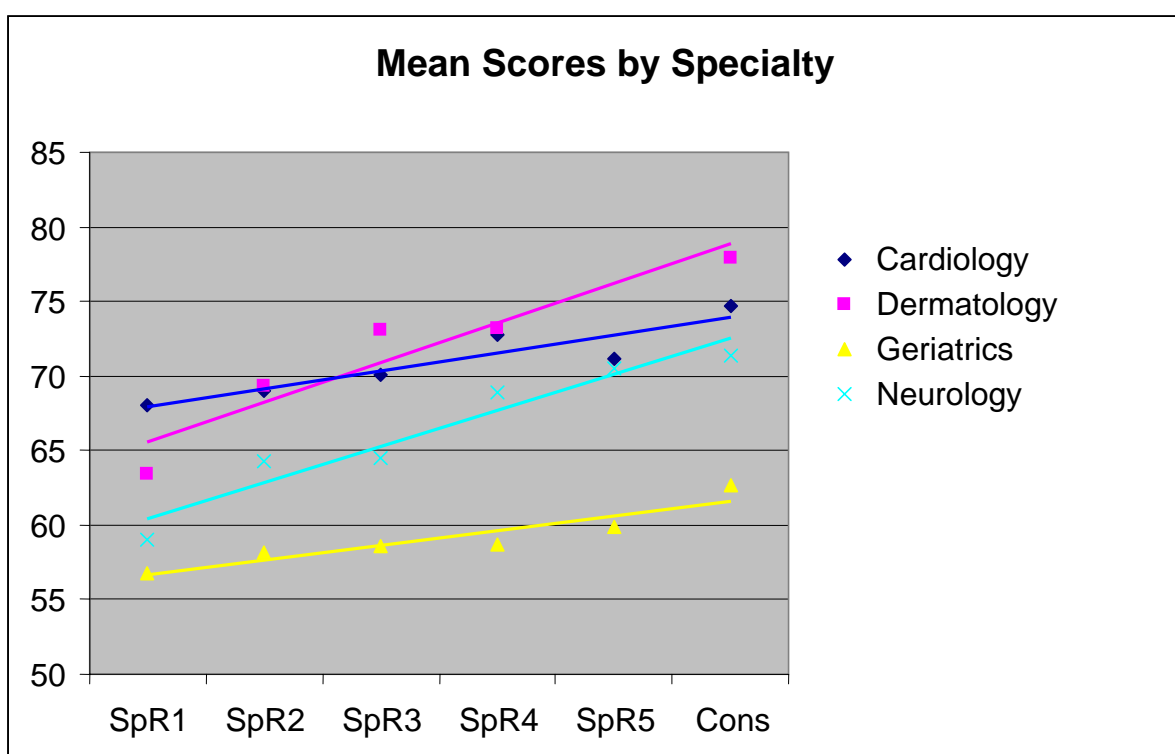
	SpRs in pilot	SpRs registered	Participation
Cardiology	280	542	52%
Dermatology	181	217	83%
Geriatrics	402	524	77%
Neurology	196	202	97%

6.2 Candidate Performance

The mean scores and pass rates of participants grouped by type of post & year of training are shown below.

6.2.1 Mean Scores by specialty/group

	SpR1	SpR2	SpR3	SpR4	SpR5	Cons	Overall
Cardiology	68.03	69.04	70.07	72.74	71.12	74.67	70.01
Dermatology	63.37	69.36	73.11	72.48	n/a	77.91	70.74
Geriatrics	56.77	58.11	58.59	58.65	59.90	62.62	58.56
Neurology	59.00	64.33	64.49	68.89	70.50	71.40	65.92



6.2.2 Pass Rates by specialty/group

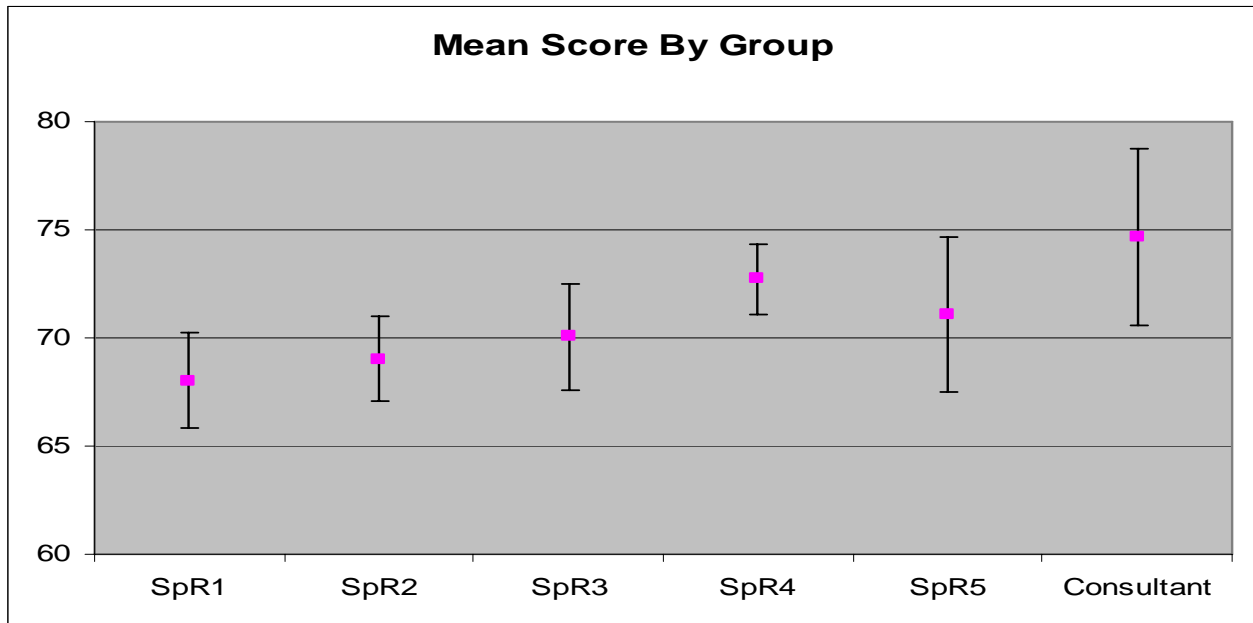
The table below shows the percentage of participants by specialty and group who would have passed the exam had the pass mark set by standard setting been applied. No pass mark for geriatric medicine was set by formal standard setting (see section 14.3 below) but a notional pass mark of 50% was applied after the paper was taken.

	SpR1	SpR2	SpR3	SpR4	SpR5	Cons	Overall
Cardiology	0.0	2.4	8.3	7.8	7.7	8.0	4.8
Dermatology	63.0	88.7	98.1	96.7	n/a	97.0	87.3
Geriatrics	83.1	88.8	91.1	96.1	91.7	97.3	87.5
Neurology	59.6	84.8	83.7	92.1	100.0	96.7	84.5

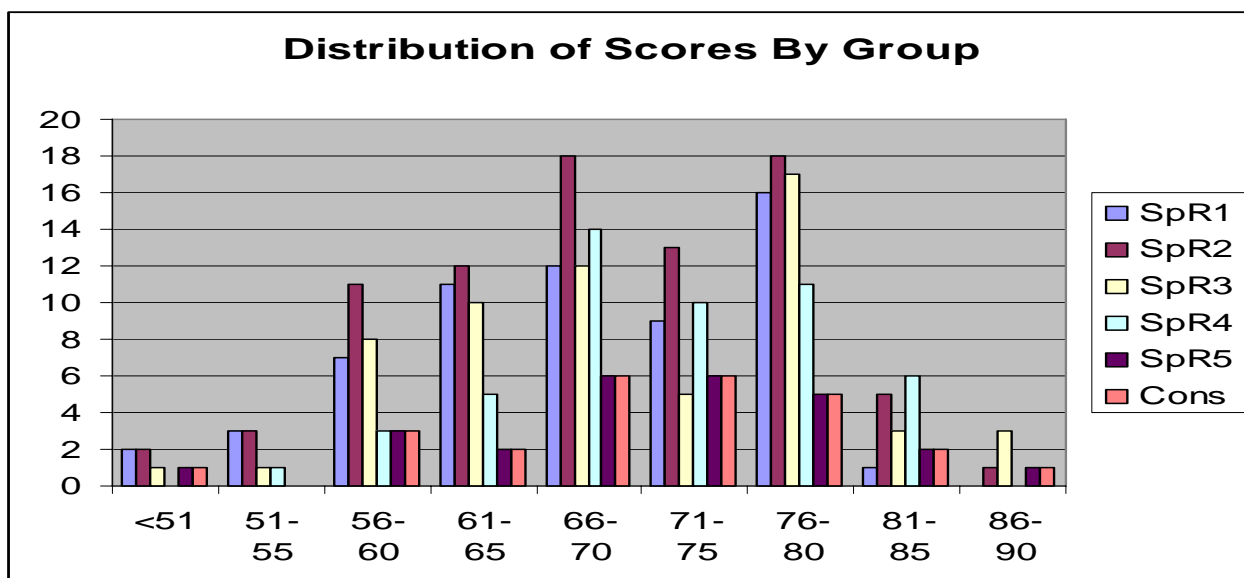
6.3 Cardiology

	N	Mean	Std. Deviation	95% Confidence Interval for Mean		Minimum	Maximum	Pass Rate
				Lower Bound	Upper Bound			
SpR1	61	68.03	8.67	65.81	70.25	48	82	0
SpR2	83	69.04	8.94	67.08	70.99	44	86	2.4
SpR3	60	70.07	9.51	67.61	72.53	42	90	8.3
SpR4	51	72.74	7.59	70.10	74.37	54	92	7.8
SpR5	25	71.12	8.68	67.54	74.70	50	86	7.7
Consultant	12	74.67	6.40	70.60	78.73	66	88	8.0
Total	292	70.01	8.77	69.00	71.02	42	92	4.8

Pass mark from standard setting: 83%



t-bars show 95% confidence intervals



Participants' Comments

77 participants wrote comments on the questionnaire sheets. Apart from specific comments on perceived deficiencies or errors in individual questions these have been broken down into the following themes:

- Will all SpRs sit at the same time? (1)
- Syllabus should be clearly defined (1)
- Is the assessment valid? (1)
- Lack of negative marking not clear (3)
- Feedback would be helpful (6)
- Should be part of the annual RITA rather than pass/fail (2)
- Images should be included (8)
- Should be more basic science (1)
- Should be less basic science (1)
- The correct answers will vary from hospital to hospital depending on local facilities (6)
- Questions on guidelines are difficult to answer because of variation (4)
- Drug questions are difficult because treatment is an art rather than a science (2)
- Questions good and relevant (25)
- Questions not always detailed enough (5)
- Scenarios generally realistic (11)
- Some ambiguous questions (14)

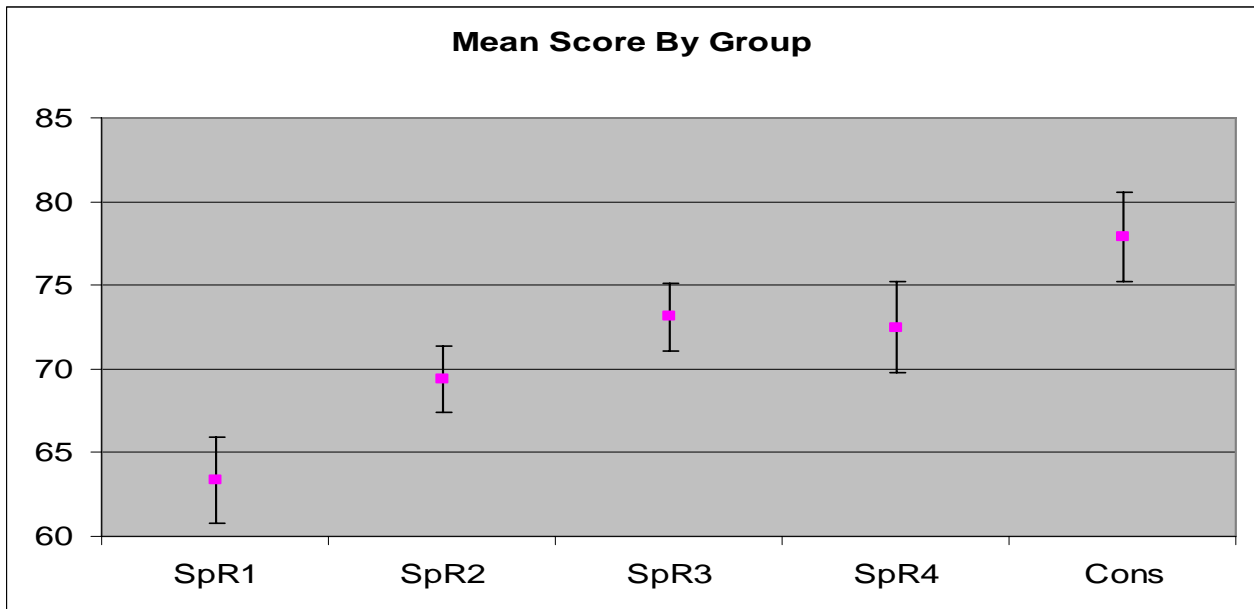
'This is an unfair assessment. I cannot see why this is necessary and I am unhappy about being asked to do it.' Second year SpR

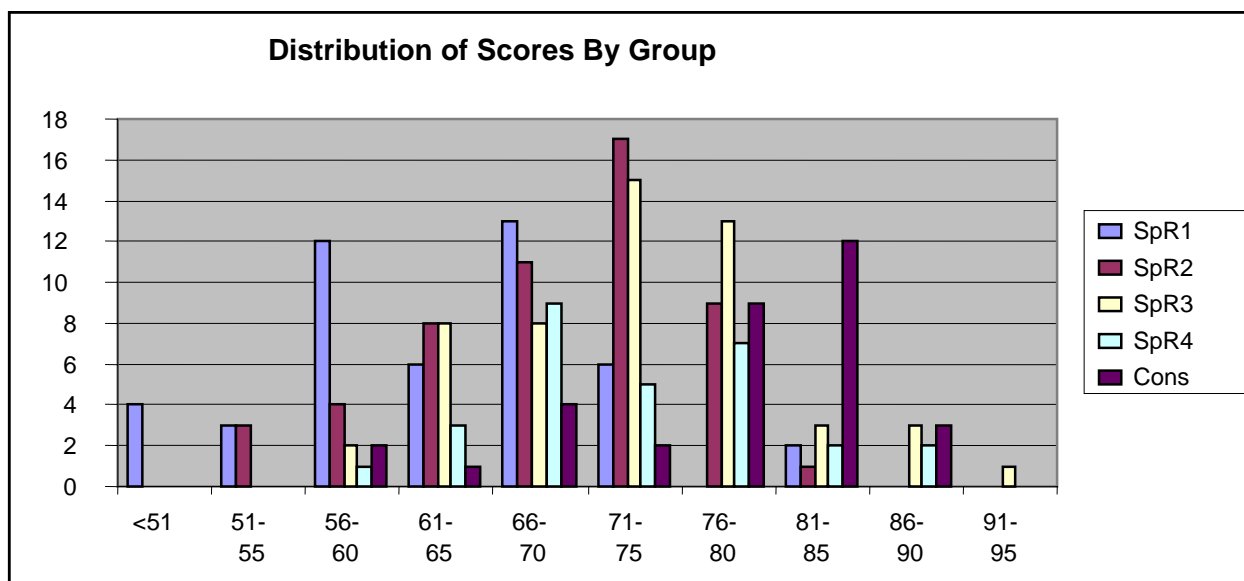
'A good set of questions which reflect everyday clinical practice. A useful stimulus to read and keep updated. Excellent method of objective assessment and should be part of RITA every year.' Third year SpR

6.4 Dermatology

	N	Mean	Std. Deviation	95% Confidence Interval for Mean		Minimum	Maximum	Pass Rate
				Lower Bound	Upper Bound			
SpR1	46	63.37	8.89	60.80	65.94	45	83	63.0
SpR2	53	69.36	7.38	67.37	71.35	52	81	88.7
SpR3	53	73.11	7.51	71.09	75.13	59	91	98.1
SpR4	29	72.48	7.45	69.77	75.19	57	88	96.6
Consultant	33	77.91	7.84	75.24	80.58	59	87	97.0
Total	214	70.74	9.07	69.53	71.96	45	91	85.1

Pass mark from standard setting: 60%





Participants' Comments

97 participants wrote comments on the questionnaire sheets. Apart from specific comments on perceived deficiencies or errors in individual questions these have been broken down into the following themes:

- Poor quality of images (66)
- Too much tropical medicine (6)
- Radiotherapy regimes not required (2)
- Some ambiguous questions (8)
- Good questions / assessment method (12)
- Opposed to this assessment method (1)
- Too many male genital questions (1)
- Too many paediatric questions (1)
- Paediatric questions difficult (2)

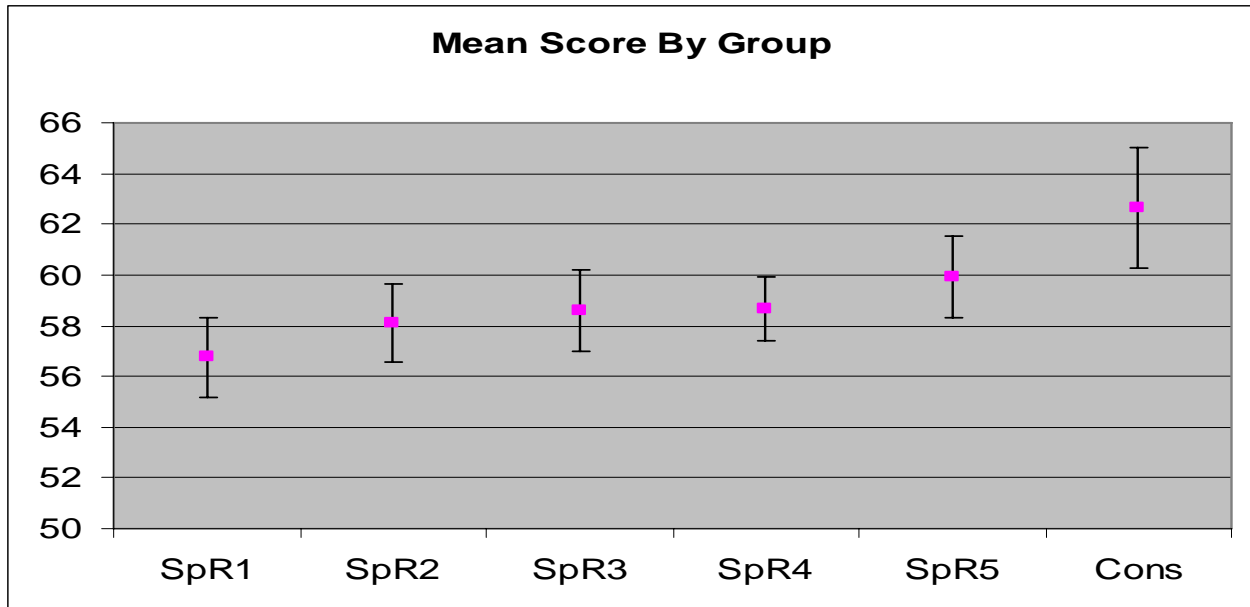
'Very useful for dermatology to have a written assessment as the subject matter is so wide.'

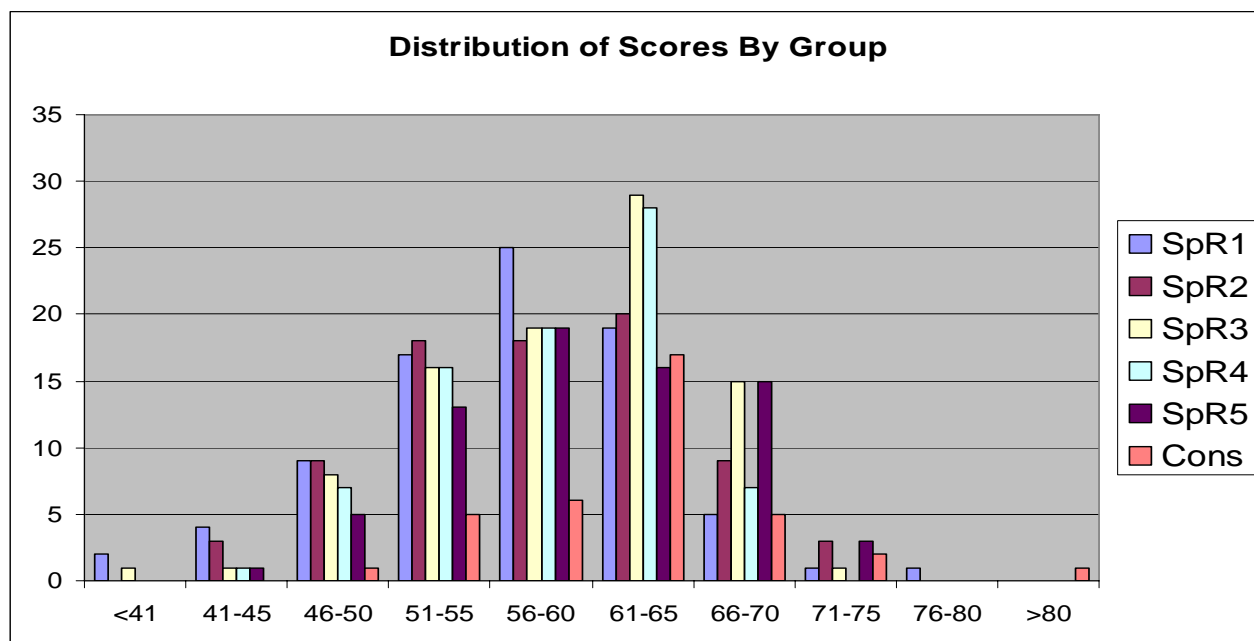
'I have thoroughly enjoyed the exam experience. It has given pointers towards areas of deficiency in knowledge and clinical practice methods.' First year SpR

6.5 Geriatric Medicine

	N	Mean	Std. Deviation	95% Confidence Interval for Mean		Minimum	Maximum	Pass Rate
				Lower Bound	Upper Bound			
SpR1	83	56.77	7.29	55.20	58.34	36	76	83.1
SpR2	79	58.11	7.07	56.55	59.67	41	73	88.8
SpR3	90	58.59	7.74	56.99	60.19	15	72	91.1
SpR4	78	58.65	5.74	57.38	59.93	43	70	96.1
SpR5	72	59.90	6.93	58.30	61.50	42	75	91.7
Consultant	37	62.62	7.40	60.24	65.01	48	90	97.3
Total	448	58.56	7.27	57.89	59.24	15	90	87.5

Post-hoc pass mark of 50% applied





Questionnaire Responses				
How difficult did you find the questions?	difficult	about right	easy	no response
	29.0%	63.8%	0.4%	6.7%
Was 3 hours the right time for the exam?	too short	about right	too long	no response
	2.5%	69.9%	21.7%	6.0%
Were you familiar with the question format?	yes	no	no response	
	79.7%	14.1%	6.3%	

Participants' Comments

254 participants wrote comments on the questionnaire sheets. Apart from specific comments on perceived deficiencies or errors in individual questions these have been broken down into the following main themes:

- Some ambiguous questions (47)
- Poorly written questions, e.g. too long, negative questions, double negatives (26)
- Good questions / paper (34)
- Too long / too many questions (28)
- Support for this assessment method (12)
- Opposed to this assessment method (17)
- Too many falls questions (8)
- Inappropriate orthopaedics questions (19)
- Inappropriate question on English law for Scots (6)

'Thank you. Enjoyed the exam. Gave me motivation to study and to learn more about Elderly Care Medicine. This MCQ exam is a useful way of assessing the SpR.' Third year SpR

'Multiple choice format is an unsatisfactory way of assessing candidates clinical knowledge'
First year SpR

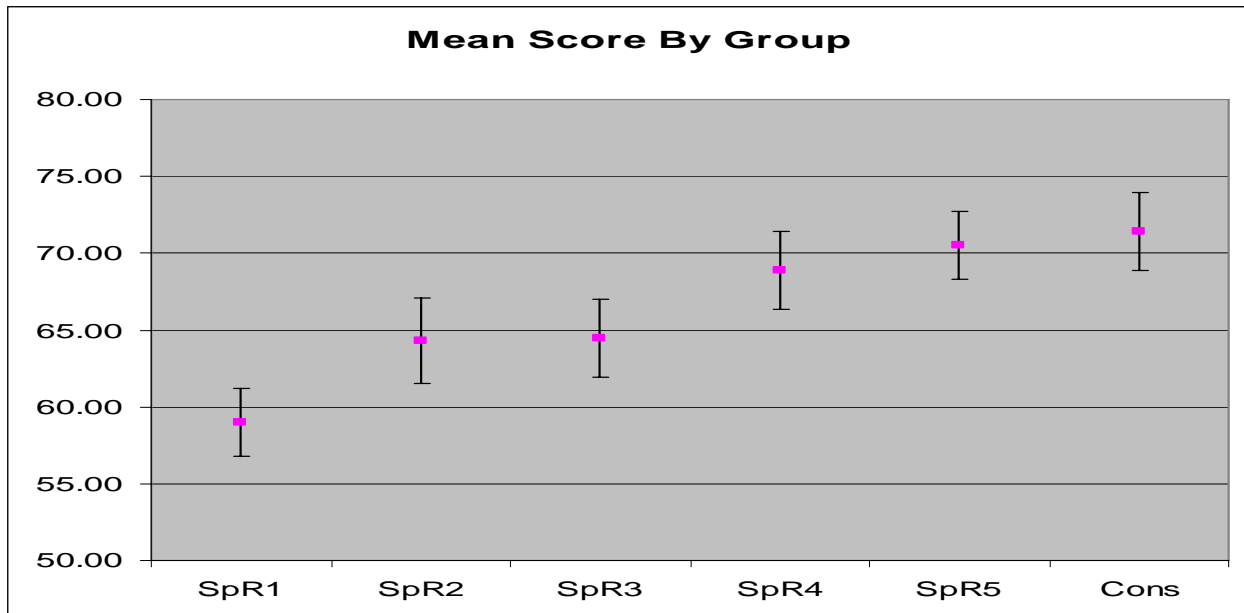
'I am not sure about outcome of these kinds of assessment. Does it make for better clinicians? Can we identify the best clinicians from scoring patterns?' Third year SpR

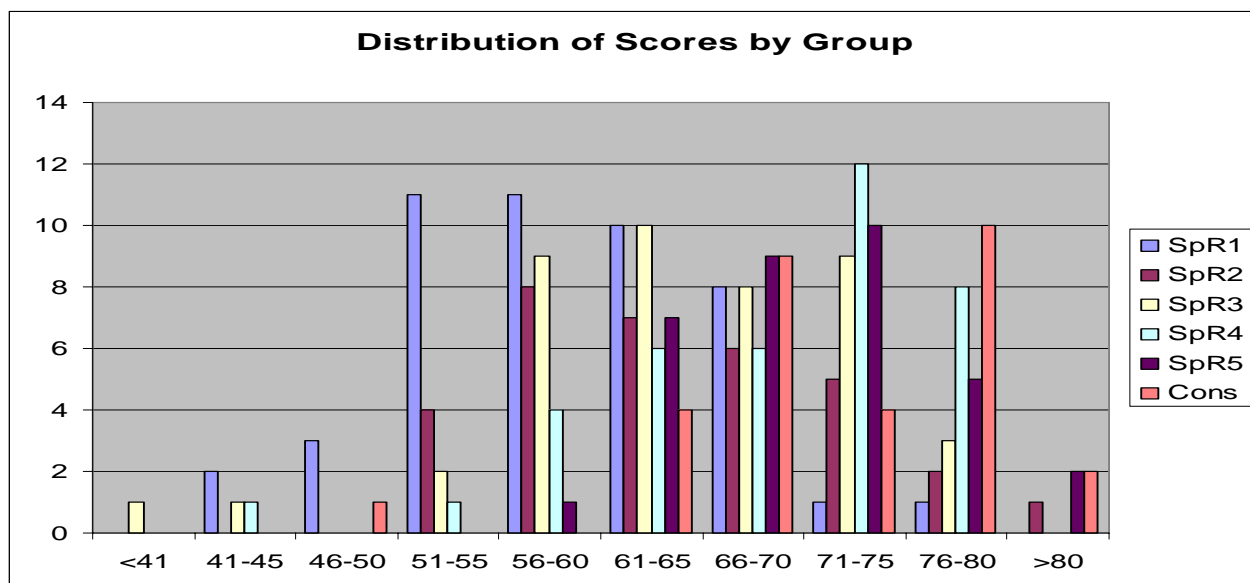
*'The questions per se are fair as long as one continually reads **factual** knowledge during a life-time of clinical practice. This is a good additional method of assessment for our training. It's a great encouragement for us to read more thoroughly.'* Fourth year SpR

6.6 Neurology

	N	Mean	Std. Deviation	95% Confidence Interval for Mean		Minimum	Maximum	Pass Rate
				Lower Bound	Upper Bound			
SpR1	47	59.00	7.61	56.82	61.18	41	77	59.6
SpR2	33	64.33	8.19	61.54	67.12	51	81	84.8
SpR3	43	64.49	8.50	61.95	67.03	40	80	83.7
SpR4	39	68.89	8.06	66.37	71.42	45	80	92.1
SpR5	34	70.50	6.53	68.30	72.70	57	83	100.0
Consultant	30	71.40	7.12	68.85	73.95	50	83	96.7
Total	226	65.92	8.83	64.77	67.07	40	83	84.5

Pass mark from standard setting: 57%





Questionnaire Responses				
How difficult did you find the questions?	difficult	about right	easy	no response
	37.0%	56.8%	0.4%	5.7%
Was 3 hours the right time for the exam?	too short	about right	too long	no response
	6.2%	74.9%	14.1%	4.8%
Were you familiar with the question format?	yes	no	no response	
	87.7%	7.9%	4.4%	

Participants' Comments

93 participants wrote comments on the questionnaire sheets. Apart from specific comments on perceived deficiencies or errors in individual questions these have been broken down into the following main themes:

- Some ambiguous questions (10)
- Good questions / paper (12)
- Too long / too many questions (23)
- Support for this assessment method (7)
- Opposed to this assessment method (5)
- Use pictures / videos (3)

'This exam has stimulated my thought process and prompted me to read up a few areas which I now recognise I need to revise.' First year SpR

'We have enough exams in our career – don't need another one. This exam does not predict how good one is clinically in everyday neurology.' Fourth year SpR

6.7 Reliability

The reliability of a multiple-choice examination can be measured using Cronbach's coefficient alpha. This is a measure of internal consistency, and it indicates how much of the information gathered is 'true' information, and how much is error. It is a number between 0 and 1, with higher values indicating a more reliable examination. High-stakes examinations are usually expected to have an alpha value exceeding 0.80, and preferably higher than 0.90.

Reliability is affected by three major factors – the number of questions, the quality of the questions and the spread of ability of candidates.

If all other factors hold true (number & ability of candidates, quality of test questions) then the reliability of a longer exam can be estimated using the Spearman-Brown formula. The table below shows the calculated reliability of the pilots (numbers in bold) and extrapolations of what the reliability of a longer examination with similar characteristics might be.

	Number of questions		
	50	100	200
Cardiology	0.55	0.71	0.83
Dermatology		0.81	0.90
Geriatrics		0.67	0.81
Neurology		0.78	0.88

7 Question Bank

All questions that have been reviewed by question setting groups are held in a database. Based on the experience of selecting and editing the question papers, it is likely that many of these will require further final editing before they can be used in an exam and some may not be usable.

The current number of questions in the bank, including those used in the pilots, is:

	Agreed	Attention Needed	To be reviewed	Rejected	Total
Cardiology	268	22	1	7	298
Dermatology	169	17	6	3	195
Geriatrics	101		27	2	130
Neurology	341	4	3	8	356

8 Discussion

8.1 Question Setting Process

The main reason for gastroenterology not setting a paper was that the approach taken failed to generate enough questions of sufficient quality. An attempt was made to delegate responsibility for groups of questions to BSG sub-sections. This did not work well – questions were slow to arrive (if they came at all) and very limited in number and quality. As with other specialties, question writers found it hard to take on the guidelines for the style required but the channels for giving feedback and requesting reworking were limited and ineffective. This outcome suggests that personal engagement with, and commitment from, individual question writers is needed. Experience with all specialties suggests that attendance at a forthcoming question setting meeting was often needed to focus authors on the task – some people did send in questions even when not attending the meetings, but more promised to do so and failed to deliver. Geriatric medicine did set a paper without the same level of commitment from question writers as the other three specialties but this was largely down to the efforts of the board, especially the chair, and would not be sustainable in the longer-term.

8.2 Question Style

It was clear that providing written guidelines to question writers was not enough to guarantee appropriate questions. Initial contributions from many writers suggested that they had either not fully read or taken in the guidelines. Training, involvement in peer review and feedback on their contributions all helped to improve quality. Central editing was still needed to ensure consistency of style. A simple example is that many questions start with a phrase in the standardised form “A 50-year-old woman presented with...”. Typically contributors might leave out the required hyphens, abbreviate year to yr, use “lady” or “female” instead of the preferred “woman”, use “attended your clinic” instead of “presented” and so on.

The process for setting the geriatrics paper involved less peer-review and editing than other specialties. The board also decided to allow some questions that were acknowledged to not meet the style guidelines. These included negatively phrased questions (e.g. “what is the least likely risk factor?”) and true-false style questions with 1 true answer (e.g. “which of the following is most true about xxx?”). These questions did not necessarily perform worse than others (though some performed badly) but it was noticeable that there were a significant number of complaints about poorly worded questions and about negative questions in particular.

8.3 Images

In general we avoided using questions with images for the pilot, to simplify the delivery of the paper.

The geriatrics papers contained one drawing embedded in the text.

The dermatology paper did use 12 images (histopathology slides and photographs) to support 7 questions. These were printed professionally on high-quality paper in a separate booklet and were checked by the question writer who had provided the histopathology slides. Despite this many dermatology participants complained that the images were of inadequate quality. Clearly this needs attention in future. *N.B the original images were not provided to the technical standard requested in the MRCP(UK) question writing manual.*

All the specialties have expressed interest in making more use of images in future - still images initially, but possibly some moving images eventually. A number of participants commented that images would help.

8.4 Attitudes to Knowledge Assessment

JCHMT had no power to compel trainees to take part, but some of the specialties did make it clear, through supervisors, that participation was “expected”. Participation rates were generally much higher than had been hoped for but it can’t be assumed that this is indicative of a positive attitude to summative assessment. All participants, however, had the opportunity to comment on the concept and in each case only a small proportion of them used this opportunity to express clear opposition.

Some comments, and anecdotal discussions during the course of the project, revealed an attitude of “it’s inevitable so we might as well get on with it”

8.5 Reliability Measures

Surprisingly high levels of internal reliability were achieved. This could be reduced in real examinations by the effects of smaller, more homogenous cohorts, though further improvements in the quality of question-setting could off-set this. It is unlikely that the alpha for real examinations will be any better than in this pilot without increasing the number of questions.

The MRCP(UK) Central Office has recently been working with psychometricians from some of the American examining boards exploring methods of calculating reliability in different circumstances. There is a view that other measures, such as analysis of standard error at the “cut score” (pass mark), may be superior to Cronbach’s alpha for this type of examination.

8.6 Number of Questions / Timing

The cardiology paper was only 50 questions in a generous time allowance of 2 hours, so raised few concerns about length, but produced an expected low reliability coefficient

The other papers were all 100 questions in 3 hours. In each case a number of participants finished and left early suggesting that the timing was about right for the length of paper. This is backed up by responses from neurology and geriatrics participants who were specifically asked about this. The question was somewhat ambiguous. For respondents who said it was too long, it is possible that they meant 100 questions were too many rather than 3 hours was unnecessary for this number.

In written comments there were complaints that it was hard to concentrate for this long on this many complex questions without a break.

If future assessments are delivered by computer, BS7988 *Code of practice for the use of IT in the delivery of assessments* contains the following clause in relation to timing:

“6.4.5.1 For assessments longer than 1.5 hours and where the candidate works almost entirely at the screen there should be provision for candidates to take a break”

8.7 Value to Question Writers

Many of the people who attended question setting, paper selection and standard setting meetings commented on what a valuable and enjoyable educational experience it was. Reviewing questions was the catalyst for much debate about, for example, the relative merits of management techniques, the evidence base to support questions, and the ethical/legal basis for decisions.

8.8 Question Writers' Time

Unfortunately we did not manage to gather any statistical data about the time spent by question writers. From anecdotal discussions we believe that much question-writing was carried out at evening and weekends, on leave, on train journeys etc. Attendance at question setting groups required a full day (with travel) for each meeting. Many writers commented that it is increasingly difficult for them to get their trusts to allow them to spend time on this type of educational activity. This may present a challenge to the setting of future examinations and it would not be wise to be over-reliant on expecting doctors to commit large amounts of their own time to these activities.

9 Conclusions

Setting high-quality specialty examinations using best-of-five questions is a feasible activity for the Federation of Royal Colleges of Physicians.

It is possible to achieve acceptable levels of reliability using this method.

There is widespread (but not universal), tacit acceptance by trainees that this form of assessment is appropriate and, indeed, inevitable.

The 5 specialties who took part in the project now have a good understanding of the complexity and scale of the task needed to set a specialty examination.

The setting and operation of high quality specialty examinations requires experience and expertise. It is most appropriate that the existing MRCP(UK) organisation should take on this responsibility.

Partnership between the Colleges and the specialist societies has been invaluable to the pilot and we should seek to extend this partnership to the setting of specialist examinations in future.

Question writers need a combination of written guidelines, training and feedback to make them effective. They need a personal commitment to, and relationship with, the process – remote connections by delegation of question-writing to sub-groups does not work

The existing MRCP(UK) standards and guidelines for question writing should be adopted for new specialty examinations.

Computerised delivery of the assessment would allow for effective delivery of images and for future use of moving images.

Specialist examinations should ideally be based on 200 questions to maximise the reliability. This will however present a challenge to question setting groups and may not be achievable immediately.

Research should be undertaken into the best methods to use to demonstrate reliability of specialty examinations, given the specific characteristics of these assessments (small cohorts, compressed range of ability). Data from this project could be used to inform that research.

The existing project boards and question setting groups are well placed to take this work forward, though in some cases they may need supplementing/reorganising to boost expertise and ensure suitable representation of stakeholders.

The question bank developed by the project should be retained for future use.

For some small specialties there may not be enough of a critical mass of consultants to set a paper or enough trainees to make it reliable.

10 Acknowledgements

This project was entirely funded by a grant from the Department of Health, for which JCHMT is extremely grateful.

Apart from those people who have already been named in the body of the report and the appendices we would like to acknowledge the following:

Support and advice from the officers and staff of the MRCP(UK) organisation has been invaluable to this project, particularly John Mucklow, Jim Benson, Kate Beaumont, Jennifer Mollon, Alyx Jenkins, Rachael O'Flynn.

Darin Nagamootoo from the RCP Medical Workforce Unit for the design of questionnaires.

Over 100 Regional Specialty Advisors, Programmer Directors, chairs of Specialty Training Committees, deanery administrators and others who committed time and effort to the local organisation and invigilation of the pilots.

Question setters are also too numerous to mention individually, but nothing could have been done without them.

Everyone who took part in the pilot assessments.

11 Appendix A - Cardiology Details

11.1 Project Board

The Joint Cardiology Assessment Board (JCAB) was established by the Cardiology SAC and the British Cardiovascular Society to coordinate the project:

Dr Tony Mourant (chair)	BCS Education Committee
Dr Frank Dunn	BCS Education Committee
Dr Khalid Bharakat	SAC
Dr Theresa McDonagh	SAC
Dr Peter Mills	SAC
Dr Gordon Murray	BCS Education Committee

There was no trainee representative on the board but the British Junior Cardiologists Association was kept informed and one of their representatives attended one JCAB meeting.

Dr Clive Lawson did the final proof-reading of the paper.

11.2 Question Setting Group

35 consultant cardiologists (including JCAB members) agreed to write questions. Two QSG meetings were held, in May and October 2005. 24 writers managed to attend at least one of these meetings. Some questions were sent in for meetings by people who did not personally attend. Since the pilot some people have been removed from the group but we have also managed to recruit others and currently have 34 potential question-writers.

The QSG included nominated representatives from the following BCS Affiliated Groups:

- British Society of Echocardiography
- British Association for Cardiac Rehabilitation
- British Society for Heart Failure
- British Cardiovascular Intervention Society
- British Congenital Cardiac Association
- British Nuclear Cardiology Society

11.3 Approach

The standard approach for question and paper setting, as described above in section 4 was largely followed. Writers were allocated specific topics to write on. Standard setting was not done as a separate process but was combined with the meeting that selected and edited questions for the final paper and some participants had not rated questions in advance.

In order to preserve questions for future use a decision was taken to limit the pilot paper to 50 questions.

11.4 Follow-On

Results were presented at a BCS Conference in Glasgow on April 24th.

A QSG meeting was held in June 2006 in an attempt to keep up the momentum on question writing. Another QSG meeting is being planned for November 2006.

Dr Mourant and Dr Mills have written a brief report with specific recommendations for cardiology.

12 Appendix B - Dermatology Details

12.1 Project Board

The British Association of Dermatologists Education Committee was asked to coordinate the project, though with contributions from the SAC.

BAD

Dr Robert Charles-Holmes (Chair), Prof David Gawkrödger, Dr David Fitzgerald, Dr Clive Archer, Prof Eugene Healy, Dr Giles Dunnill, Dr Ian Coulson, Prof James Ferguson, Dr Richard Groves, Dr Saleem Taibjee (SpR representative), Dr Stephen Jones, Dr Sue Burge, Dr Jane McGregor and Dr Graham Ogg

Dr Mike Tidman did the final proof-reading of the paper.

SAC

Dr Chris Bunker

Discussion and decision-making was largely undertaken in smaller sub-groups or by e-mail.

12.2 Question Setting Group

30 consultant dermatologists and 1 SpR (the trainee rep) agreed to write questions. Two QSG meetings were held, in June and November 2005. 27 writers managed to attend at least one of these meetings. A few questions were sent in for meetings by people who did not personally attend.

12.3 Approach

The standard approach for question and paper setting, as described above in section 4 was largely followed. Writers were generally allocated specific topics of special interest to write on. Questions that were accepted during the QSG meetings were put through a second level of review by a subject expert after the meeting. Standard setting was combined with the meeting that selected questions for the final paper, but did follow quite a rigorous process.

12.4 Follow-On

A poster was presented at the BAD annual meeting in July 2006. Questions that were not used in the pilot are currently being reviewed. No further question-writing is currently planned.

13 Appendix C - Gastroenterology Details

13.1 Project Board

Dr Ian Forgacs, SAC member and chair of the BGS Training Committee, was asked to coordinate the project and was responsible for most decision making with the SAC chair, Dr Bob Walt. The project was discussed at two meetings of the BGS Training Committee attended by the project manager.

13.2 Question Setting Group

An attempt was made to generate questions by delegating responsibility for groups of questions to BSG sub-sections (10 sections, 20 questions each). Section leaders were asked by the President of the BSG to take part and they were given written guidance on question style. This did not work well – questions were slow to arrive (if they came at all) and very limited in number. As with other specialties, question writers found it hard to take on the guidelines for the style required but the channels for giving feedback and requesting reworking were limited and ineffective.

A QSG meeting was held in December 2005 at which 8 attendees reviewed and edited some of these questions. This group of 8 (which included some with MRCP experience) could form a core for taking forward future gastroenterology question setting.

13.3 Withdrawal

In March 2006 Dr Forgacs and Dr Walt agreed to suspend all activity on question setting because the lack of progress with question writing meant that any pilot assessment would run much later than other specialties. By then it seemed likely that a new MRCP specialty exam would proceed and they decided that they would wait to be in the first wave of that process.

14 Appendix D - Geriatric Medicine Details

14.1 Project Board

A project board was established to represent the SAC, the British Geriatrics Society and the Diploma of Geriatric Medicine.

Membership:

Prof Steve Allen (chair)	SAC
Dr John Gladman	SAC
Dr Oliver Corrado	BGS / SAC
Dr Kevin Kelleher	BGS / SAC
Prof Stuart Parker	BGS / DGM
Dr Chris Turnbull.	SAC

Dr Michael Vassallo did the final proof-reading of the paper.

14.2 Question Setting Group

No real question setting group was established. Board members wrote many questions themselves and approached colleagues and contacts to write some. A total of 12 writers contributed questions. One question-writing training session was held, attended by board members and other question writers. The project manager did some editing for matters of style and format and Prof Allen reviewed and edited all questions himself.

14.3 Approach

A paper-selection meeting was attended by all board members and one other question writer. Questions were reviewed and edited at this meeting. Final replacement and rewriting of rejected questions was carried out via e-mail discussion after this meeting.

No standard setting was carried out to set a notional pass mark. The project board thought it would be a false science to set a notional pass mark for the pilot on the grounds that geriatric medicine is very broad and knowledge thresholds are difficult to define from a standing start. Moreover, ambiguity is an inevitable aspect of geriatric medicine, so it can't be easily stripped out of exams without over-simplification; this will continue to be a problem in the future.

14.4 Follow-On

The results were presented to the SAC in June 2006. No further question setting is currently planned, but the SAC has decided in principle to continue to build up the question bank, under the leadership of Dr Oliver Corrado.

15 Appendix E - Neurology Details

15.1 *Project Board*

The Neurology Training and Assessment Committee (NTAC) was created by the Neurology SAC in partnership with the Training and Education Committee of the Association of British Neurologists to oversee development of assessment techniques for trainee neurologists. NTAC's overall aim is to improve the quality of neurology training through the development and implementation of appropriate assessment techniques. NTAC took on the role of the project board for this project.

NTAC Membership:

Chair of TESC - ex officio

Chair of SAC - ex officio

2 members neurology SAC (nominated by SAC)

2 members of TESC (nominated by TESC)

2 trainees, members of ABN(T)

RCP Education Department representative

RCP Project Manager

1 neurology programme director

1 educational supervisor

neurophysiology observer

Co-opted (for expertise in question writing and knowledge assessment)

Prof Mark Wiles

Dr Geraint Fuller

Dr Fred Schon

Dr Gavin Young

Dr Richard Davenport

Dr Adrian Wills

Dr Camille Carroll / Dr Claire Hirst

Dr Connie Tengah

Winnie Wade

Joe Booth

<vacant>

Dr Lionel Ginsberg

Dr Adrian Fowle

Dr John Scadding

15.2 *Question Setting Group*

Approximately 40 consultant neurologists and 2 SpRs (the trainee reps) expressed interest in writing questions. Three QSG meetings were held, in July, October and December 2006. 19 writers managed to attend at least one of these meetings. A good number of questions were sent in for meetings by people who did not personally attend (often via colleagues who were attending). In a few cases local question setting sessions were held to feed into the central QSG meetings. In all we ended up with questions contributed from over 30 named writers.

15.3 *Approach*

The standard approach described above for question and paper setting was followed. Writers were generally not allocated specific topics to write on, though topics where the question bank was weak were highlighted before the final QSG meeting.

During the paper selection and final editing stage questions were edited with intense attention to detail to fit with the MRCP(UK) writing style.

Standard setting was done by a group that, with the exception of the chair, was completely separate from the group that selected the paper, and included an SpR and a new consultant.

15.4 *Follow-On*

A short presentation was given at the ABN conference in October. Performance of individual questions is being reviewed in detail. A QSG meeting is planned for October 2006.