Preface to the HPAC 10-year review and the RCGP response to the recommendations

In 2017 the Trustee Board of the Royal College of General Practitioners (RCGP) commissioned an external review of the MRCGP examination in recognition that it had been running for 10 years as the licensing exam for General Practice. Following a tendering process, Health Professional Assessment Consultancy (HPAC) were asked to undertake this review.

Their report, below, was completed in September 2017 and the RCGP were delighted to read the reviewers found that overall the CSA and AKT “meet or exceed the standards for procedures used for high stakes examinations in the medical profession” and that the CSA and AKT “were fit for purpose and fair for both candidates and patients”. As one would expect, they did make a series of recommendations “in the spirit of continuous quality improvement”.

The RCGP exams team have reviewed and considered the implications of the recommendations in consultation with a stakeholder group including representation from Health Education England (HEE), the devolved nations, Associates in Training, patients, First5s (GPs in the first five years after qualification), Trainers and Educators who had the opportunity to question and advise the exams team about the proposed responses.

The aim of this document is to make the report public, in line with RCGP policy on exam matters, and in doing so to include the RCGP response to the recommendations. RCGP comments have been embedded within text boxes in the original report in the appropriate places for ease of reading. Most of the RCGP response is embedded in the executive summary, however there are a few recommendations in the main body of the report only and our response to these can be found in the same place.

MRCGP Examinations team
25 October 2018

6 October 2017

Final Report from Health Professional Assessment Consultancy on the 10-year external review of the MRCGP examination

Executive Summary of Final Report for the 10-Year MRCGP Review conducted by HPAC in 2017

Introduction

This is the Executive Summary of our final report, in which we summarise the work conducted in reviewing the MRCGP, relate the findings to the Kane Validity Framework, national guidelines, best evidence assessment practice and offer some suggestions for improvements.

In the spirit of continuous quality improvement, we have made some suggestions for potential enhancements to the MRCGP procedures.

Process

The review was conducted between April and June 2017. This included

- Documentary analysis (electronic on RCGP website, by access to restricted documents from the RCGP and by email from various members of the MRCGP officers) using **Kane Validity Framework (KVF) – see Appendix 1**
- Personal and teleconference interviews with several relevant MRCGP post holders to enhance documentary evidence and explore actual practice and implementation
- Attendance and observation of AKT Committee activities (May 2017), applying the KVF
- Attendance and observation of a CSA at the RCGP (May 2017) applying the KVF
- Psychometric analyses of AKT and CSA test forms over the past 10 years (specialised psychometric focus)
- A specific focus on the MRCGP with respect to fairness, both to candidates and patients, and priority areas for change to improve fairness

Findings and suggestions

General overview

Overall, the reviewers were impressed by the availability of information on the MRCGP website, which was extensive and transparent, with ample information for potential candidates.

The Applied Knowledge Test (AKT)

The statement of purpose of the AKT is clearly articulated and is an excellent example of the first section of the Kane Validity Model, i.e. the intended purpose of the test. (see Appendix 1 for the Kane Validity Framework)

The Applied Knowledge Test (AKT) is a summative assessment of the knowledge base that underpins independent general practice within the context of the UK National Health Service. Candidates who pass the AKT have demonstrated their competence in applying knowledge at a level sufficiently high for independent practice.

(www.rcgp.org.uk)

Overall, the AKT meets or exceeds standards for procedures used for high-stakes examinations in the medical profession.

Recommendations related to the AKT

1. Increase the percentage of items presented as clinical vignettes and provide more detailed patient information in each vignette.

RGCP proposed action:
Comment on recommendation
Any increase in item length risks adding more pressure on all candidates in a time-limited exam and there is a risk of a detrimental effect on candidates with Specific Learning Difficulties and those with English as a second language.
Action to be taken
To partially implement this recommendation in the first instance by increasing the length of clinical vignettes in those very short items with reduced item performance statistics only, but not for very short items which are known to perform well.
Plan to review
To monitor whether such edits improve or reduce subsequent item performance, the effects on the mean and median time taken including, if feasible, for different candidate subgroups over a minimum 1-year period after implementation.

2. Rather than drawing the common-items equating block from a single test form, assemble it from multiple test forms.

RCGP proposed action:
Comment on recommendation
Over the past ten years, the use of anchor-items from multiple test forms has been successfully used, but not on a regular basis as is now being recommended. Implementation should enhance the ability of the RCGP to demonstrate that the standard required to enter general practice is being maintained.
Action to be taken
To implement this recommendation in full by using anchor-items from multiple test forms, the technical detail of which will be based on further psychometric advice.
Plan to review
Review standards of differing exams 1-year after implementation

3. As a security precaution, randomize the order in which items are presented to candidates.

RCGP proposed action:
Comment on recommendation
We have no evidence that security is breached in this module and the AKT is already randomised. With the suggested change, candidates will no longer get a standardised experience, some starting with 'harder' items than others, some having a short run of items all about the same clinical topic, risking a potential for low level cueing of the correct answer. There are significant technical and contractual issues with the test centre provider related to implementation.
Action to be taken
Pilot an alternative randomisation method of a sample exam using test-centre software with input from our new psychometric team.
Plan to Review
After the pilot has been run and evaluated we will review the risks and benefits of implementing for real.

4. Make the following (minor) modifications in item analysis and key validation procedures.
    a. In addition to calculating item statistics based on the total group of candidates, also calculate and retain item statistics based upon graduates of UK schools taking the AKT for the first time.
    b. Rather than point-biserial correlations, use corrected item-total biserial correlations as the index of item discrimination.
    c. Calculate the Horst statistic and apply it in identification of items for review by subject matter experts during key validation.

RCGP proposed action:
Comment on recommendation
These are technical recommendations which relate principally to test content and construction. They have no direct impact on candidates taking the AKT.
Proposed action
Recommendations 4a) and 4c) to be included as part of the standard psychometric procedures.
Recommendation 4b) will require piloting and further psychometric advice before full implementation.
Plan to review
This will happen routinely during psychometric review of the exam statistics after each diet

5. Increase the amount of time allotted per item without decreasing the total number of items.

RCGP proposed Action:
This is a combined response to the recommendations numbered 5 and 6 in the main body

Comment on recommendation 5 - **Increase the amount of time allotted per item**
To increase the amount of time per item without increasing test length, means that the number of items would need to be reduced. Doing this would risk reducing test reliability when the AKT currently performs at very high levels of internationally accepted measurements of reliability. Curriculum coverage would be reduced which would impact on the other modules especially the WPBA, at a time when there is a national drive to reduce the burden of assessment in WPBA.
Action to be taken
As we have evidence that almost all candidates complete the AKT within the current time allowed, the technical and financial consequences of increasing the time per question or the total test time outweigh any advantage and we will not be implementing this recommendation.
Plan to review
To continue to monitor the metrics of the time taken by candidates throughout the AKT and review if recommended by RCGP psychometricians.

Comment on recommendation 6 - **Increase the amount of time allotted per item without decreasing the total numbers of items**
Essentially this recommendation can only be achieved by increasing the time allowed for the whole exam, currently 190 minutes. The number of candidates omitting items is consistently very low and increasing the standard exam time by 10 minutes in 2014 (after a previous external review recommendation) had minimal impact.
Additional test time has significant procedural difficulties in relation to administering the exam twice in one day and would inevitably mean the higher burden of cost being passed onto candidates. Any extension to current test time means that all candidates would require a formal rest break to reduce fatigue and therefore, for test security, the delivery of two separate tests before and after a coffee break.
Proposed action
As we have no evidence from our data that candidates need longer to complete the AKT (those who have specific learning difficulties are allowed additional time) and due to the potential adverse effects of increased fees and extended test time on candidates we will not be implementing this recommendation.
Plan to review
We will continue to review the data after each exam relating to time taken per item and omission rates for individual items.

6. Enhance score reports to provide candidates with information about potential areas of strength and weakness by increasing the number of content areas reported and including mean percent-correct scores for a reference group (probably UK graduates taking the AKT for the first time) for each content area.

RCGP proposed action:

Comment on recommendation

Giving more detailed feedback on an increased number of content areas is problematic as too much detail could compromise test security and the ability to reuse items, with significant cost implications and consequent increased fees to the candidates. Scores for subsections with very small numbers of questions will give unreliable and potentially misleading feedback. Published evidence and the clear view of the RCGP's independent psychometric advisers is that it is potentially harmful to provide such information.

After discussion the stakeholder group felt that detailed feedback benchmarked against UK graduates could emphasise differential attainment in an unhelpful way. They felt that giving mean cohort scores for each of the three AKT sub-domains (clinical medicine, data interpretation and GP administration), in addition to the mean total cohort score which is current practice, would go some way to increasing meaningful and useful feedback given to candidates in a more granular way.

Proposed action

To give enhanced feedback to candidates by incorporating the mean cohort scores for each of the three domains with their own results. Changes to the ePortfolio may need to be accomplished before this can be implemented.

Plan to Review

I year after implementation by candidate and trainer survey.

The Clinical Skills Assessment (CSA)

The statement of purpose of the CSA is clearly articulated and is another excellent example of the first stage of the Kane Validity Model, i.e. the intended purpose of the test. (see Appendix 1 for the Kane Validity Framework)

The aim of the CSA is to test a doctor's ability to gather information and apply learned understanding of disease processes and person-centred care appropriately in a standardized context, make evidence-based decisions and communicate effectively with patients and colleagues.

(www.rcgp.org.uk)

Overall, the CSA meets or exceeds standards for procedures used for high-stakes examinations in the medical profession.

Recommendations related to the CSA

1. Use generalizability theory to analyze the reproducibility of scores.

RCGP proposed action:
Comment on recommendation
This recommendation relates to the statistical analysis used to quality assure the CSA exam and suggests use of an alternative method. The newer generalizability approach is becoming more widely used, though more complex and expensive, and is accepted as giving more accurate estimates of the reliability and accuracy of this type of exam. We used generalizability theory to analyse the CSA in 2009 and the findings were congruent with the methods to estimate reliability we were then using and continue to use today.
As the exam runs over a period of eight months of each year, it would be most appropriate to analyse data over a year
Planned action
To commission a generalizability study using one year's collated data, then consider the results to identify realised advantages with a view to making a final decision on any change.
Plan to Review
Review data from pilot generalizability study in approximately one year

2. Increase the number of stations in the CSA to improve the reproducibility of scores.

RCGP proposed action:
Comment on recommendation
We currently sample the curriculum with a palette of 13 cases in purpose-built premises. This enables us to run two exams a day over three circuits (72 candidates). The initial CSA pilot trialled 16 stations and this was reduced to 13 due to significant fatigue in candidates, role-players and examiners. Since then we have regularly considered the possibility of increasing the number of stations to increase reproducibility and each time decided not to due to issues of potential fatigue, the cost implications and logistical difficulties. Having considered the issue of test length again in the light of this report, the gain in reproducibility is unlikely to be significant unless more than four cases are added and that then introduces major logistical difficulties. The stakeholder group was agreed that we should not change test length at the current time.
Planned action
Not to increase the number of stations at present
Plan to Review
Review the decision with new data which will become available after a Generalizability study. Such data would more accurately assess the number of stations needed to significantly impact reliability.

3. Set the passing standards using borderline regression methods.

RCGP proposed action:
Comment on recommendation
The CSA currently uses the Borderline Group (BG) method for standard setting. This method was considered best practice when it was adopted in 2010 to replace the previous 'Number Needed to Pass' system. Since then, there has been significant development in the field of psychometrics, and the recommended Borderline Regression (BR) method is now considered to be the gold standard. The exam team has considered changing to the BR method and modelled the likely impact on a number of occasions prior to this review. Then, as now, modelling suggests that there would be a downward effect on the pass rate if we were to change to the BR method. As the BG method remains mainstream, reliable and fit for purpose, after full and detailed discussions of the pros and cons of initiating this change the RCGP has decided not to do so at the present time.
Planned action
To continue using the BG method for standard setting
Plan to Review
The RCGP will keep standard setting under active review taking account of current changes to selection and training; for example, if the GMC were to express a preference for us to use the BR method at some stage in the future this would precipitate a reassessment of the current method.

4. Conduct regression analyses to identify stations/examiners that might demonstrate aberrant characteristics.

RCGP proposed action:
Comment on recommendation
The CSA already uses an extremely robust system of quality assurance of both cases and examiners, based on performance metrics and exception reporting, leading to tight monitoring and possible re-writing of cases or re-training of examiners. This is far more sophisticated than systems in use in similar exams. At the request of the Stakeholder group, we obtained further psychometric advice regarding a change from this to regression analyses. Neither the longstanding psychometrician nor the new team perceived an advantage.
Planned action
To continue to use the current methods of quality assurance
Plan to review
To review this decision in the light of the generalizability study.

5. Explore the use of key-feature style checklist items in combination with global rating scales for station scoring to decrease variability in marker stringency and increase consistency in the marking criteria used.

RCGP proposed action:
Comment on recommendation
Key-feature style checklist items are shorter and more general. They have been shown to decrease the cognitive load of examiners and are better attuned to the global rating scales that we use. The RCGP had previously agreed to move in this direction at some stage and are happy to explore it now.
Planned action
Phased adoption of key-features style check list in marking schedules. This will happen gradually as part of the case QA system and examiners will be trained in the use of the modified marking schedules.
Plan to review
As part of regular exam psychometric review to ascertain whether the change has a beneficial effect on variability in marker stringency.

Fairness

Overall, the reviewers thought that the CSA and AKT were fit for purpose and fair for both candidates and patients. We have made some recommendations motivated by the potential to increase fairness.

Summary

In relation to the Kane Validity Framework, our views are as follows:

The first stage of the Kane Validity Framework (KVF), i.e. the statement of purpose of the examinations, is clearly and explicitly stated, for the overall programme of the MRCGP assessments, as well as for each component.

The second stage of the KVF – the five domains of validity evidence: the reviewers were able to find all the evidence required for to evaluate each of the domains, and it was clear that the RCGP had considered the various aspects in relation to designing and implementing a programme of assessment which met international standards.

Stage 3, the Interpretive Argument (how all the components explained in the evidence section come together to form a strong case for using the assessment, from a validity perspective) is also more than adequately addressed.

Overall, the team of reviewers considered that the Applied Knowledge Test (AKT) and the Clinical Skills Assessment (CSA) components of the MRCGP met, or in some places, exceeded standards for procedures used for high-stakes examinations in the medical profession. This view was unanimously agreed by all the reviewers in relation to both national and international perspectives, based on their extensive experience of current best practice and the medical education literature. We also considered that MRCGP in its current form (and with suggested developments), is in line with the GMC's SCAR recommendations and Generic Professional Capabilities requirements.

Main Final Report

Introduction

The 10-year Review of the MRCGP was commissioned by the RCGP Trustee Board, as the 'new' format was introduced in 2007, and a review was considered desirable.

This is our final report, which summarises the work conducted, evaluates the findings and offers some suggestions for improvements.

We used the Kane Validity Framework (Appendix 1) to establish how we evaluated the overall programme of examinations as well as the AKT and CSA components individually.

We also evaluated the MRCGP in relation to the GMC's SCAR (Standards for Curriculum and Assessment Review) recommendations and General Professional Capabilities requirements.

The reviewers evaluated the examinations with particular reference to fairness for candidates and patients.

In the spirit of continuous quality improvement, for potential enhancements to the MRCGP procedures, we have made some suggestions specific to the AKT and CSA.

Process

The review was conducted between April and June 2017. This included

- Documentary analysis (electronic on RCGP website, by access to restricted documents from the RCGP and by email from various members of the MRCGP officers) using Kane Validity Framework (KVF)
- Personal and teleconference interviews with several relevant MRCGP post holders to enhance documentary evidence and explore actual practice and implementation
- Attendance and observation of AKT Committee activities (May 2017), applying the KVF
- Attendance and observation of a CSA at the RCGP (May 2017), applying the KVF
- Psychometric analyses of AKT and CSA test forms over the past 10 years (specialised psychometric focus)

We would like to comment that we found a high degree of co-operation from the RCGP staff, in enabling the reviewers access to documentation, arranging for attendance at events and setting up interviews with the relevant individuals as requested.

The review process investigated if the theory and practice of assessment, using Kane's model of validity, was followed. Kane's three stages relating to 1) 'intended purpose/use', 2) 'meaningful evidence', and 3) 'argument: justifying the decisions' were followed under this framework.

The reviewers also examined the evidence in relation to GMC's SCAR (Standards for Curriculum and Assessment Review) recommendations and General Professional Capabilities requirements.

Findings and suggestions

General overview

Overall, we were impressed by the availability of information on the MRCGP website. The first stage of the Kane Validity Framework (KVF), i.e. the statement of purpose of the programme of examinations, is clearly and explicitly stated. Stage 3, the Interpretive Argument (how all the components explained in the evidence section come together to form a strong case for using the assessment, from a validity perspective) is also more than adequately addressed.

We reviewed the evidence relating to Stage 2 of the KVF in the sections about the AKT and CSA.

Fairness

Overall, the reviewers thought that the CSA and AKT were fit for purpose and fair for both candidates and patients. We have made some recommendations motivated by the potential to increase fairness.

Fairness to Patients: This depends heavily on ensuring the accuracy and reproducibility of pass/fail decisions to ensure that passing candidates are able to provide safe and effective patient care.
The recommendations related to increasing the reproducibility of scores and shifting to the use of borderline regression for standard setting in the CSA are offered for this reason.
We also feel strongly that the current procedure of adjusting the pass/fail standard upwards, using the standard error of measurement, is highly appropriate: from the standpoint of protecting patients, protecting against “false passer” decisions (passing a candidate who does not merit it) is more important that making “false failer” decisions (failing a candidate who actually meets the standard), particularly given that candidates have an opportunity to repeat examinations.

Fairness to Candidates:
Differences in passing rates for domestic and international medical graduates are commonly observed on certification examinations for other UK Royal Colleges and in many other countries at licensing and certification level.

The reviewers found no characteristics of the examination design or test administration procedures which would cause the differences seen.
The reviewers considered that the examination outcomes reflect true differences in the knowledge and skills required to practice safely and effectively.

Component-specific review

Applied Knowledge Test (AKT)

The Applied Knowledge Test (AKT) is a summative assessment of the knowledge base that underpins independent general practice within the context of the UK National Health Service. Candidates who pass the AKT have demonstrated their competence in applying knowledge at a level sufficiently high for independent practice.

The subsections are as follows:
Subsection 1A describes the procedures used to develop, administer, analyze, score, and report scores on the AKT.
Subsection 1B provides a series of recommendations for potential enhancements to those procedures.

1A Overview of Procedures Used for the AKT

Item Development and Review

There is detailed guidance available for question selection, see document “AKT selection instructions 2017”

The guidance for item writers is that items developed must be relevant, topical, either common conditions or high impact. There are 10 item writers at any one time. There is no fixed term. There is an application process for those who want to become item writers (examiners).

The essential attributes of an item writer (examiner) are explicitly and clearly stated:

- Demonstration of a GP knowledge base adequate for contemporary clinical practice – a pass in the AKT within the last 10 years. Applicants who sat the MRCGP MCQ prior to 2006 would be able to take the current AKT to demonstrate their clinical knowledge base.
- A certain natural aptitude for the task as shown by submitted questions
- A completer-finisher type personality who follows detailed instructions and submits work within deadlines
- Appropriate levels of IT literacy
- A degree of numeracy which is required for some of the psychometric concepts

- An interest in the development of the AKT and the Essential Knowledge Challenge (EKC)
- Ability to work in a small group and contribute effectively
- An ability to accept feedback and criticism, and to feel comfortable exposing gaps in your knowledge

It is preferable that examiners are GP trainers, must have worked for at least five years as a GP and are not beyond two years after retirement.

Those applying are asked to fill in an application form and submit five single best answer questions. They are sent suitable information on how to write a question as well as the house style - based on the NBME style guide (see “AKT question bank formatting template”). These questions and then scored anonymously and individually by members of the AKT core group. The application forms themselves are also scored by the core group.

Item Review

When a question is selected, if there have been any changes in the relevant clinical guidelines since its last use, that person is responsible for updating the question and sending it (securely, see later) to other core group members for review. There is not a particular person or group in charge of keeping the exam bank up to date but rather updating is performed on questions as they are to be used. If a guideline is noted to have changed, a question can be flagged as requiring updating the next time it is due to be used. There is an audit trail for this with version control and responsibility for this task and for the final sign-off for each AKT.

Following collation of marks, at a committee meeting comprising the psychometrician, administrator and the core group, results are scrutinised at length.

Questions specifically discussed by the group are those with a facility under 40% and those with a point biserial under 0.20. In addition, anchor questions previously used were examined where there was a significant difference in results to a different sitting of that question, as this may reflect a change in clinical guidelines, making it unsuitable to use as an anchor in the future.

Poorly performing questions may be removed from the exam, incorrectly keyed answers (in the correct answer grid) may be corrected, or marks awarded for more than one answer where scrutiny reveals a question to be ambiguous or where there could be more than one best answer. Incorrect text answers are

manually examined in Excel format to ensure these can be added to the answer key (e.g. the use of a comma instead of a decimal point by European candidates). The number of new questions in each assessment is noted at the meeting, this is not set but seems to vary between 30 and 50 for the most part. New questions are not pre-tested – the group have in fact published on this to show no difference in performance or pretested and non-pretested questions:

Dixon H, Blow C, Milne P, Siriwardena N. Reliability of non-pretested versus pretested questions in the applied knowledge test (AKT) of the MRCGP: evidence of quality assurance. Educ Prim Care. 2014;25(3):149-54. https://www.ncbi.nlm.nih.gov/pubmed/25198471

They have also undertaken an external linguistics review of the language used as well as an external assessment of bias towards non-UK/Irish candidates. Both favourable.

Item Formats. Four item formats are used on the AKT: single-best-answer MCQs (some including images and graphics), extended matching questions, completion of tables/algorithms, and (small numbers of) free-text answers. The majority of the items consist of brief clinical vignettes describing a patient care situation and asking the candidate to demonstrate an understanding of the situation by (e.g.) indicating a diagnosis or the next step in patient care. Irrespective of the question format, candidates are awarded one mark for each item answered correctly, and marks are not deducted for incorrect responses or for failing to respond.

Test Construction. Three forms of the AKT are assembled and administered each year. Each form consists of 200 items built according to a test blueprint designed to sample across the general practice curriculum. On each test form, 80% of the items are on clinical medicine, 10% on research/evidence-based practice, and 10% on legal/ethical/administration issues. To support common items equating (see below), 40 items are taken from the test form given one year earlier. Items are selected so that, taken together, they are content-representative of a full test paper, and care is taken to avoid items that are very easy and very hard and those that test “emerging knowledge.” Items selected for the equating block also have relatively high discrimination indices (item-total point-biserial correlations greater than 0.20). Occasionally, blocks of common items from multiple previous test forms are used. In addition to the common items used for scaling and equating, a number of additional items have been used on previous test forms; for security reasons, items appearing from test forms used in the two immediately preceding administrations are avoided.

Test Administration. The AKT is generally taken in the second or third year of training in general practice. It is a 190-minute computer-based exam administered three times annually (January, April and October) in morning and afternoon sessions at Pearson Vue professional testing centres around the UK. The same test form is used with a quarantine period for candidates at the changeover time for test security. Professional invigilators are used with CCTV monitoring. For each administration, roughly 1000 candidates take the AKT for the first time and about 300 more repeat it. In 2015-16, roughly 85% of first-time test takers were UK graduates, 3% were EEA graduates, and the remaining 12% graduated from medical schools elsewhere in the world. Candidates receiving test accommodations requiring extra time take the AKT in the afternoon at Pearson Centres open for extended hours. For each administration, items are presented in the same (fixed) order for all candidates.

Test and Item Analysis, Key Validation, and the Reproducibility of Scores. After each AKT administration, a preliminary scoring and item analysis are performed, producing standard item statistics for each item, including a facility index (p-value), a discrimination index (point-biserial correlation coefficient), and the percentage of candidates selecting each response. The item statistics are used to identify items for "key validation" – review by a committee of subject matter experts. This review determines if aberrant statistics have resulted from incorrect answer keys, and a decision is made on whether or not to include each of the flagged items in calculation of scores. All 200 items are scored and only rarely is an item deleted.
After key validation, a final scoring is conducted, and indices of the reproducibility of scores (coefficient alpha, the standard error of measurement – SEM) are calculated. For recent administration, the median value of the estimates of coefficient alpha has been 0.90, and the median SEM has been 2.8%.[1] In addition, an analysis is conducted to determine if there are "runs" of omitted items at the end of the test; these have typically shown that the vast majority of candidates respond to all of the items.

Scaling and Equating. A common-items linear equating is done using performance of the full candidature on the 40 items drawn from the test form administered 12 months previously. A "reality check" using data on reference

[1] Coefficient alpha can be interpreted as the expected correlation between candidates' scores on test forms covering similar content with different (randomly parallel) items. For high-stakes tests, a value 0.90 is highly satisfactory. The SEM indicates the extent to which scores can be expected to vary by chance; it can be used to form confidence intervals around scores. For example, if the SEM is 2.8% and a candidate's score is 70%, there is a 68% probability that candidate's "true" score is between 67.2% and 72.8% (the score +/- 1 SEM), indicating that observed scores are quite reproducible (precise).

groups is also conducted to verify the accuracy of the equating. (Occasionally, a more complex "double anchoring" design is used to cross-check consistency of the pass mark, between different cohorts e.g. 15 months as well as 12 months apart). The scores are not actually scaled; instead, the common-items equating is used to locate the score (standard) to be used in making pass/fail decisions.

Standard Setting. In addition to locating the pass mark statistically for each test form, the RCGP uses a modified Angoff procedure at least every three years to ensure the standard remains fit for purpose. This procedure requires review of each item on a test form by a group of standard setters that includes RCGP examiners, trainers, trainees, newly qualified GPs, representatives from the Deaneries and the British Medical Association, lay and patient members. Each standard setter predicts the percentage of just-passing candidates that would answer the item correctly. These values are averaged across standard setters and items to determine an overall pass/fail standard, which is then adjusted upward by one standard error of measurement to set the passing mark.

Pass/Fail Rates. In recent years, the pass rate for first-time candidates has been 75% to 80%. It is higher for UK graduates (85% to 88%) than for EEA graduates (56% to 59%) and graduates from the rest of the world (49% to 54%). Regardless of the location of the medical school, women candidates pass the AKT at a rate several percentage points higher than men. As one might expect, pass rates are highest on first attempts and decrease on subsequent attempts. Across groups, the “ultimate pass rate” (across up to five attempts) for recent cohorts is 96% to 98%, so the vast majority of candidates eventually pass the AKT.

Score Reporting. Candidate score reports include the candidate name; the date of the test administration; the candidate’s pass/fail result; overall percent-correct score; percent-correct marks for clinical medicine, evidence interpretation, and organizational questions; the percent-correct score required to pass; and the mean overall mark for the cohort taking the AKT on that test date.

Annual Report on Examination Performance. For almost a decade, the RCGP has published an annual report (available at http://www.rcgp.org.uk/training-exams/mrcgp-exams-overview/mrcgp-annual-reports.aspx) summarizing AKT and CSA outcomes. These reports include overall pass/fail outcomes for the associated year; performance as a function of candidates’ protected characteristics (sex, primary medical qualifications, ethnicity, disability); selected performance trends; and results of a wealth of well-conceived additional analyses. The RCGP is to be congratulated for its transparency in providing such detailed information about candidate performance. Review of that information was very helpful in preparing this report.

1B Recommendations related to the AKT

Overall, the AKT meets or exceeds standards for procedures used for high-stakes examinations in the medical profession. In the spirit of continuous quality improvement, this section offers some recommendations for potential enhancements to those procedures.

1. ***Increase the percentage of items presented as clinical vignettes and provide more detailed patient information in each vignette.***
   Patient-based MCQs are, in effect, low-fidelity clinical simulations that challenge candidates to interpret patient findings and decide on the next step in patient care. When seeing patients, GPs have a wealth of information available at a glance: patient age and gender, body habitus, how sick the patient looks, etc. Even for familiar patients, the GP will discuss the reason for the visit, take a history of the present illness, and conduct at least a brief physical exam. Providing this kind of information in a paragraph-length stem is desirable.

2. ***Rather than drawing the common-items equating block from a single test form, assemble it from multiple test forms.***
   There was some evidence in the reports that candidate performance is higher for October administrations. While this is may well be due to "seasonality" in when different types of candidates choose to take the AKT, an alternative explanation is that the pass/fail standards for the January, April, and October administrations have drifted apart as a result of drawing the items used in equating from a single administration taking place a year earlier. An alternate approach is to systematically assemble multiple common-item blocks (e.g. draw 25 items each from the AKT28, AKT29, and AKT30 test forms for use in equating AKT33). In this example, each of the three common-item blocks would provide a basis for estimating the pass mark for AKT33, and these could be averaged to determine the pass mark actually used. (Operationally, three common-item equating blocks could be set aside after each administration for use on later administrations.) This should provide a more precise basis for estimating the pass mark, and if the pass/fail standards have drifted apart over time, this should eventually bring them back together.

3. ***As a security precaution, randomize the order in which items are presented to candidates.***
   Currently, items are presented in the same (fixed) sequence to all candidates. While it seems fairer to standardize the item order, it also poses unnecessary

security risks. For example, a group of candidates could agree in advance who would memorize item 1, 2, 3, etc, making it easy to reconstruct test content after the administration.

Alternatively, a candidate could leave a list of answers (1C, 2A, 3B, 4D, 5A, 6D) in a WC stall for another candidate to retrieve. If there is concern that randomization may adversely affect the performance of some candidates (e.g. because items requiring more time may appear earlier in a test form, resulting in less time to complete items near the end), a better solution for coping with the potential effects of inadvertent "differential speededness" is to simply allow a little more time per item.

4. ***Make the following (minor) modifications in item analysis and key validation procedures.***

   a. In addition to calculating item statistics based on the total group of candidates, also calculate and retain item statistics based upon graduates of UK schools taking the AKT for the first time.
   Though AKT sample sizes are fairly large, there are small shifts in the proportion of test-takers from different candidate groups from administration to administration, and these can affect item statistics. Also archiving statistics for UK first-takers should ameliorate this problem to some degree.

   b. Rather than point-biserial correlations, use corrected item-total biserial correlations as the index of item discrimination.
   Point biserial correlations tend to be somewhat unstable across candidate groups differing in proficiency, particularly for easy items. Biserial correlations are more stable and are easy to estimate from point-biserials. Regardless, discrimination indices should be calculated as corrected item-total correlations (the item score should be omitted from the total score in the calculation); it was unclear from the available documentation how the calculation was done.

   c. Calculate the Horst statistic and apply it in identification of items for review by subject matter experts during key validation.
   The Horst statistic is the difference between the percentage selecting the correct answer and the most popular distractor. It is very useful in identifying items that may have a second (or no) correct answer for review during key validation. (As an example, regardless of the value of an item discrimination index, an item with a facility index (p-value) of 0.48 and a Horst index -0.04 is much more likely to have a second correct answer than an item with a Horst index of 0.30.

d. During key validation, consider omitting review of items with high facility indices (p-values) and mildly negative or small positive discrimination indices.

Items with facility indices greater than 0.70 are very unlikely to have incorrect answer keys or have second correct answers, regardless of the value of the discrimination index. (In part, this reflects the imprecision of estimates of item discrimination, even with fairly large sample sizes.) It is not worth spending the time of subject matter experts in review of such items. The Horst index will provide a better basis for flagging items for review that may have been mis-keyed or have second correct answers.

RCGP Comment

This was a misunderstanding at the time of the review and this does not feature in the HPAC Executive Summary recommendations list. HPAC later withdrew this recommendation when they understood our processes fully.

5. *Increase the amount of time allotted per item.*

Currently, candidates have less than one-minute per item, which is faster pacing than most exams in which items are predominantly in a clinical vignette format. While analyses looking at omit-rates indicate that the vast majority of candidates respond to all items, this is not a particularly sensitive indicator of speededness when candidates are well aware that there is no penalty for guessing (not having a guessing penalty is best practice). A more sensitive indicator of speededness is to investigate response times as a function of item-position and determine the percentage of candidates in various groups (particularly non-native English speakers) responding very quickly (e.g. in less than 10 seconds).[2] Regardless, it is appropriate for the high-stakes AKT to be primarily a "power" (as opposed to "speeded") exam, particularly given lower pass rates for those with English as a second language.

6. *Increase the time allotted per item without decreasing the total number of items.*

The time allotted per item can be increased by either decreasing the total number of items without adjusting the testing time or by increasing the testing time without changing the number of items. The latter is preferable because decreasing the number of items would decrease test reliability and increase

[2] Another analytic approach is to compare facility indices and response times for re-used items that have shifted in item position from early in the test form to late in the test form or vice versa.

the standard error of measurement. It seems desirable to re-explore the times at which centres open and close with Pearson Vue to see if it is possible to increase the time allotted for AKT administration. Ideally, the amount of the increase in time would be guided by analysis work, candidate responses to a post-test survey[3] immediately after AKT administration inquiring how much additional time would have been useful (with "no increase in time is necessary" as an option), or both.

7. ***Enhance score reports to provide candidates with information about potential areas of strength and weakness by increasing the number of content areas reported and including mean percent-correct scores for a reference group (probably UK grads taking the AKT for the first time) for each content area.***

   Though the primary purpose of the AKT is clearly to support summative decision making regarding candidates' qualifications for independent general practice, it is also desirable to provide formative ("assessment for learning") feedback to guide future learning, particularly for failing candidates. Most items can be classified along multiple dimensions (e.g. patient age and gender; organ system; clinical tasks like prevention, diagnosis, use of diagnostic studies, treatment), and these can be used for calculation of subscores for additional content areas.[4]

   To aid in interpretation, it would also be useful to provide candidates with a better sense of how well they performed relative to others (e.g. a score of 70% correct may be 10% better or worse than peers depending upon the difficulty of the associated set of items). Though it may require substantial effort at start-up, it should be straightforward to enhance score reports to incorporate such information. If the RCGP decides to move in this direction they may wish to explore use of a graphical format that also conveys information about the score precision.

[3] If the RCGP does not currently ask candidates to complete a brief post-test survey, this is generally a good idea to solicit reactions to and suggestions for AKT registration procedures, content coverage, etc.

[4] There are some strong arguments against the provision of detailed performance feedback when subscores are based on small numbers of items. The reproducibility of such scores can be quite poor; as a consequence, identified areas of strength and weakness for a candidate might change substantially if a different sample of items had been used on the test form. There are statistical (Bayesian) procedures available to address this problem by adjusting reported subscores based upon performance in other content area, though these are not commonly used.

Clinical Skills Assessment (CSA)

The stated aim of the CSA is 'to test a doctor's ability to gather information and apply learned understanding of disease processes and person-centered care appropriately in a standardized context, make evidence-based decisions and communicate effectively with patients and colleagues' (www.rcgp.org.uk).

The subsections are as follows:
Subsection 2A describes the procedures used to develop, administer, analyze, score, and report scores on the CSA.
Subsection 2B provides a series of recommendations for potential enhancements to those procedures.

2A Overview of Procedures Used for the CSA

Test Development and Construction

The CSA is a 13-station OSCE, comprising 10-minute stations that represent consultations in typical NHS general practice. Unlike the majority of OSCEs, in the CSA the candidate remains stationary in a consulting room and the simulated patient and assessor rotate around the circuit. The aim of this is to mimic a real-life general practice surgery.

The CSA is constructed in the following way:

- Cases are stored in the item bank mapped to areas of the curriculum and coded with key factors related to the case.
- A 'palette' of cases is put together for each day of testing by a dedicated member of staff following a number of criteria to ensure adequate sampling. These criteria include selection of cases based on age, gender, social class of simulated patients; cases that include a diversity element; cases that include medicines management; and cases that call for clinical examination.
- The palette is reviewed by a group of six checkers, formed from the case writing group, who are familiar with the content of cases. They cross-check for any duplication of content and appropriateness of selected cases and suggest appropriate substitutions if needed.
- The item level psychometric data is also reviewed to ensure there is variation in difficulty between cases and standardization in difficulty between diets.
- There is a final check across the **blueprint** for the whole MRCGP to ensure adequate sampling from all learning outcomes across the curriculum. On the whole, the blueprinting process appears thorough and considered.

- There is a process for reviewing the item bank and mapping cases to areas of the curriculum that are not well covered. Until recently the case writing group functioned fairly independently. There has been a move to link the case writing group with the overall strategy for the CSA through appointment of a role that sits across groups. This will allow better mapping of cases to overall curriculum and allow more targeted development of relevant cases.

- A new station has been developed recently that is a written format item, testing aspects of prescribing practice, in line with curriculum developments. This is being piloted currently.

- **There is a rigorous process for selection and appointment of case writers**. New writers are recruited through advertisement to current CSA examiners and invited to apply through submission of a test case. These cases are scrutinized and some applicants invited to further training. At this, suitable writers are invited to become part of the case writing group. Applicants are all practicing UK GPs, familiar with both the curriculum and the standard of UK training. They participate in group case writing days where partially constructed cases are brought to a face-to-face event and further developed, with the aid of role players to pilot them. Group discussion and feedback leads to further refinements. Finalized cases are then piloted by GP trainees at pilot events. Feedback is sought from trainees, role players and assessors at these events and refined to be ready for inclusion in the bank.

- Once they are in use, feedback from assessors and role players is sought after each diet and any necessary refinements made to cases. These changes are tracked and logged along with the station. Cases are reviewed prior to use and relevant updates are evident in the case log e.g. change of HbA1c units from percentage to mmol/mol in line with change in practice nationally, addition of new NICE guidance on atrial fibrillation. It is also noted that stations include reference to possible regional variations across the UK, including different commissioning arrangements and legal guidance e.g. funding for non-medical circumcision.

- The cases are kept in a secure bank in the College, each with an identification number and a key identifying relevant features for the blueprint. However, noted that while stations are being constructed they are kept on individual writer's personal computers and passed between the group via email or Dropbox, posing a potential security risk.

- Noted from the station list, there is a mix of genders and ages, with representation of paediatric cases and elderly patients. However, the names of patients are rather limited and seem to be mainly Anglo-Saxon. It could be

a security risk that makes it simpler for candidates to reconstruct the case bank over time.

- We would encourage more use of diversity in naming, and in ethnic make-up, across the blueprint to be more representative of the UK population. This may cause some difficulties in role player recruitment, but this could be managed.

- As mentioned, the entire assessment consists of a 'surgery' of 13 x 10-minute consultations. Therefore, total testing time for each candidate is 130 minutes; with two-minute intervals between each of the cases and a mid-point break, the duration of the whole examination is nearer to three hours. This offers face validity as it mimics general practice in the UK, with 10-minute appointments attended by patients who come to the doctor's consulting room. Reliability data for 13 stations has not been seen in this review, but the number has been chosen primarily based on feasibility. The current pilot for a 14th station may alter this.

- Using global rating scales, candidates are marked in each station in three domains: data gathering; clinical management; interpersonal skills. It is possible to obtain a score of 0, 1, 2 or 3 in each domain, with a maximum of 9 marks per station. The domains are chosen to reflect the broad GP curriculum. There are both generic and specific anchor statements available to assessors for marking in each domain.

Test Administration

All assessments take place in a dedicated examination centre. The day consists of two separate candidate groups undergoing assessment, one in the morning and one in the afternoon, with identical stations used. Three circuits are run simultaneously each session. Each circuit has an assessor and a simulated patient role player for each station who stay in the same stations for the whole day, assessing 26 candidates in total. Each circuit also has a marshal – a senior examiner – and an administrative floor manager who are able to troubleshoot and respond to unanticipated problems.

- Candidate briefing: candidates in each group are briefed by the senior marshal using a PowerPoint that is available to candidates online prior to the event. The briefing is encouraging and supportive while being clear about rules and regulations and the code of conduct. This is also available to candidates online. Candidates are asked to leave mobile phones with marshals before the assessment and collect them on leaving. Nothing can be taken into the exam other than a pre-agreed doctor's kit. Candidates are kept separated, with the afternoon group arriving before the morning group leave.

- Assessor briefing: this is given at the start of the day and consists of a PowerPoint and video clip for discussion. Following this, assessors move into groups of three to discuss individual stations in more detail in a calibration exercise.
- Role player briefing: the senior marshal briefs all role players, examines where relevant and ensures there are no last minute problems. All role players are experienced actors with an agency.
- Calibration: during a 90-minute exercise, a structured proforma is used to guide calibration, with one assessor nominated as the calibration facilitator. After assessor discussion, role players join and there is further discussion and a run-through of the station by all role players with assessors playing the candidate. Agreement is reached on how the station will be dealt with. This is a very thorough way of ensuring standardization.
- Stations and timing: candidates are led to individual rooms that are set up like a GP consulting room. They have 10 minutes prior to the start of the assessment to review patient notes. At the start of the assessment, the assessor checks the candidate GMC number on the door with the details on their marksheet. At the start of each station, the assessor and simulated patient enter the room. They leave either at the end of the 10 minutes, or sooner if the doctor finishes the consultation early. There is a two-minute break between stations where assessors can submit their scores and candidates can review the next set of patient notes. All timing is controlled by an electronic timer controlled by a marshal on each circuit. Each room has a linked clock in it for candidates to see. There is a 15-minute break after seven stations. The assessors and role players are kept separate from candidates. Candidates are allowed to mix, with a marshal supervising them to prevent them discussing stations or accessing other materials.
- Electronic format: candidates access patient notes via electronic tablets that are preloaded and in the rooms. Candidates can apply in advance for special dispensation due to a specific learning disability. In this case, they will be provided with paper notes. Assessors also read stations and mark on tablets. Paper copies are available in case of loss of technology.
- Interactions: If examination findings are required, the assessor has instructions to verbalise these, but otherwise there is no interaction between assessor and candidate. Occasionally, the simulated patient will carry a picture of an examination finding.
- Assessor tablets are connected via Wi-Fi to a central server to allow real-time upload of scores. They are unable to submit scores until the two-minute break to allow candidates the full 10 minutes to perform. The uploads are monitored in real time by an administrator. In the two minutes between stations, assessors and role players are directed to only confer on points of

fact. However, it was observed that several pairs of examiners and role-players discussed the candidate they had just seen, between stations, with the role-players sometimes offering opinions on performances. Currently, all scoring is done by the assessor and there is no provision for a lay score or simulated patient score, although we understand pilots have been undertaken.

Assessor selection and training

There is good evidence that this is rigorous. There is open advertisement for new assessors dependent on need. All assessors are practicing GPs who completed training more than five years previously. The College recognizes the need for diversity in assessors and aims that the assessor group should represent the UK population of General Practitioners. Applicants need to sit the AKT – the MRCGP knowledge test – and pass this to proceed to training. They then undertake a day of face to face training where they are assessed on their ability to perform exercises relevant to the assessor role. Successful candidates are then invited to be assessors.

- Annual training, including equality and diversity training, takes place at the MRCGP Conference over two days. Assessors are then expected to commit to assess at least 10 times per year. At each diet, four assessors are peer reviewed and their marking assessed. Assessors are given feedback every 18 months on their scores and performance in relation to other assessors and persistent outliers are removed.
- Inter-rater reliability: on the day attended, there was a mix of candidates of different ages, genders, ethnicities and special circumstances. Analysis of the entire MRCGP cohort demonstrates that candidates with special circumstances tend to score lower. However, we understand that there have been occasions when, for ease of administration, candidates with special circumstances have been grouped together. Psychometric analysis has apparently been performed within this cohort and demonstrated little difference in scores when controlling for age, gender and country of primary medical qualification. However, I would have a concern regarding assessor internal calibration in a group where the overall standard might be expected to be lower. We would suggest it would be preferable to ensure candidates are not grouped in any way based on shared characteristics.
- Feedback to candidates: assessors have a number of qualitative statements on the tablet for each mark sheet and are invited to select up to four points of feedback for each candidate. Candidates then receive these for each station, along with their station score. However, the statements are quite brief and generic and do not allow for appropriate feedback in some areas

e.g. data gathering. Assessors expressed dissatisfaction with these and were observed sometimes struggling to find a meaningful statement. Feedback to guide further learning may be better achieved by exploring options for more meaningful feedback.

Test/Station Analysis and the Reproducibility of Scores. A series of test and station indices are calculated for each day of test administration. Coefficient alpha is calculated for all circuits combined and for each circuit individually; the standard error of measurement is also computed. Station statistics include corrected station-total correlations and the values of coefficient alpha if each station were deleted from scoring; these are also calculated for all three circuits combined and for each circuit individually. When a low value of coefficient alpha occurs for all circuits combined, the more detailed information is reviewed to determine the reason for the low value. For recent administrations the mean value of coefficient alpha has been in the low 0.70s; while this value is similar to those for other high-stakes OSCEs, it is well below the value of 0.80 desirable for high-stakes exams. The mean value of the standard error of measurement has been around 4.7% in recent years. Because of the small numbers of candidates tested each day, considerable variation has been observed in the values for coefficient alpha, with a range of 0.55 to 0.82 reported for 2015-16 test administrations. Considerable variation has also been observed in the values of corrected station-total correlations, both across stations and for the same station across circuits, also reflecting the small numbers of candidates tested on each test date and circuit.

Scoring, Scaling and Equating. As noted above, marks on each of the 13 stations can range from 0 to 9, resulting in total scores with a theoretical range of 0 to 117. Total scores from different circuits and dates of test administration are not scaled or equated. Instead, a pass/fail standard is set for each day of test administration using the borderline groups method. Pass/fail decisions are based on whether a candidate's total mark was above or below a pass mark calculated as the pass/fail standard plus 1.64 times the standard error of measurement[5].

Pass/Fail Rates. In recent years, the pass rate for first-time candidates has been 82% to 84%. It is higher for UK graduates (89% to 92%) than for EEA graduates (60% to 66%) and graduates from the rest of the world (37% to 47%).

[5] Upward adjustment of the pass mark in this fashion seems appropriate, given the role of the CSA in protection of the public. However, as discussed below, use of coefficient alpha to estimate the SEM is questionable; a better estimate of the SEM (probably larger in magnitude) can be obtained using generalizability theory.

As for the AKT, women candidates pass the CSA at a rate higher than men. regardless of the location of the medical school.

Standard Setting. The borderline groups method is used: this was considered as the 'state-of-the-art' method for OSCEs when the CSA was first introduced. Assessors are asked to allocate each candidate to either pass, borderline or fail at each station. Calibration for this is addressed at assessor training and in the briefing on the day. If there are less than four borderline candidates for the station then a compromise passing score of 4.5/9 is allocated. If candidate numbers are small then the borderline score is taken from the day when the palette was used on a full 78 candidates, unaffected by the candidates on the smaller diet.

2B Recommendations related to the CSA

Overall, the CSA meets or exceeds standards for procedures used for high-stakes assessment of clinical skills in the health professions. In the spirit of continuous quality improvement, this section offers some recommendations for potential enhancements to those procedures.

1. ***Use generalizability theory to analyze the reproducibility of scores.*** Coefficient alpha is not an appropriate index of the reliability of scores for several reasons. First, the estimates are very unstable because of small sample sizes. True score variance is poorly estimated with the sample size tested any given day, and this will result in the large day-to-day fluctuation in the estimates that RCGP has observed. Second, there are multiple sources of measurement error present that are not accurately reflected in coefficient alpha: day-to-day differences in overall station difficulty, circuit-to-circuit and station-to-station differences in rater stringency, variation in portrayal by role players playing the same role, and case specificity. Third, the approach in current use ignores assignment of candidates to circuits and the confounding of true score variance and test form (circuit) difficulty that is present. As a result, coefficient alpha will generally result in over-estimates of the reliability of scores and under-estimates of the standard error of measurement – the precision of scores is poorer than the observed values of coefficient alpha indicate. It is highly desirable to use a generalizability theory framework to obtain more accurate estimates of the reproducibility of scores and the

resulting variance components will provide better guidance for test design and improvement[6].

IRT is unlikely to be of much value and would add substantially to complexity. Use of generalizability theory is absolutely necessary – coefficient alpha provides an inflated estimate of the actual reliability and an underestimate of the actual SEM. But this does not need to be done for every test administration date prior to score reporting. Conducting generalizability need only be done every 6 to 12 months, pooling information from all administrations during the period. The resulting estimate of the SEM can be used to adjust the pass/fail standard for administrations taking place for the next block of time.

2. ***Increase the number of stations on the CSA to improve the reproducibility of scores.***

In general, reliability coefficients of at least 0.80 are desirable for high-stakes assessments like the CSA. The median value of coefficient alpha for recent CSA administrations is significantly lower than this, and the actual reliability is lower than that estimated using coefficient alpha. We would suggest increasing the test length to roughly 20 stations[7]. Based upon a brief discussion with the RCGP psychometrician regarding the space used for test administration, it appeared that this could be accomplished by running longer morning and afternoon sessions with two concurrent circuits in each session. This would allow roughly 80 candidates to be tested per day. This is similar to the number tested currently, but with more testing time per candidate (and a heavier workload for individual examiners).

As a comparison, clinical skills tests in the US (USMLE and ECFMG) are substantially longer than three hours (as are most surgeries). Depending upon the amounts paid to role players and assessors, the increase in cost should be fairly modest. If the test length were changed to 19 stations, a morning and an afternoon session would still be possible, and 38 candidates could be tested in each session, accommodating similar numbers to the current exam. It would definitely be a longer day for candidates and examiners, however, but it seems likely that actual costs are driven more by travel-related expenses which should be relatively unaffected.

---

[6] See Swanson & van der Vleuten (2013) Clinical skills assessment: State of the art revisited in *Teaching and Learning in Medicine* for a more extended discussion of problems with coefficient alpha and the need for generalizability analyses to obtain more appropriate estimates of the reproducibility of scores

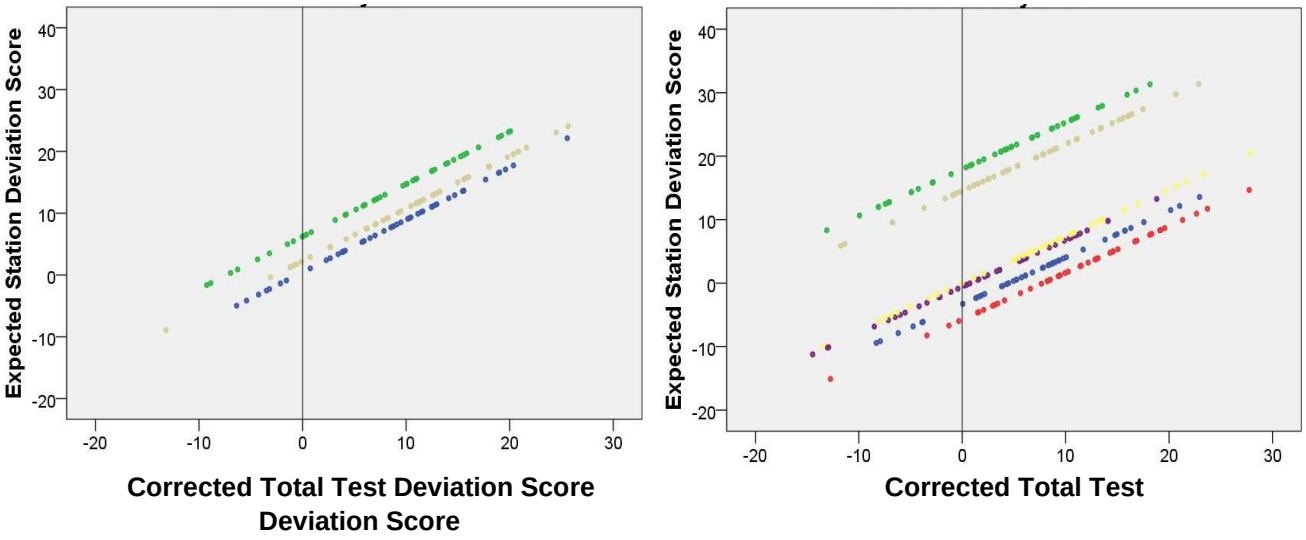
[7] The suggested increase in test length is based primarily on logistical considerations. It would be desirable to conduct the kinds of generalizability analyses described in Swanson & van der Vleuten (2013) to verify that the suggested increase is sufficient.

Sequential testing is a possibility, with a screening examination of 8 to 10 stations given initially with a follow-up exam used for those who do not clearly pass (or fail, though it seems likely to be politically difficult to fail a candidate based on a short exam). Also, many (perhaps the majority) of those who will be asked to return to complete the exam are likely to be from schools outside the UK. From a logistics perspective, this approach may result in a need for more testing dates, so the cost differences may be small, but is worthy of consideration – concentrating resources on those near the pass/fail point makes sense.

3. ***Set standards using borderline regression methods.***
For stations in which the vast majority of candidates receive an overall judgment of passing-level performance, estimated standards for individual stations are imprecise because of the small sample of candidates on which they are based. The borderline regression approach uses more of the available data to estimate standards for individual stations. If desired, a variation on the borderline regression procedure could also produce examiner-specific standards that take into account both overall station difficulty and variation in the stringency of examiners marking the same station; because this should improve the reproducibility of scores and pass/fail decisions, adoption of this approach merits exploration.
This recommendation is intended to improve the comparability of pass/fail standards across test dates, which is not currently reflected in the SEM as it is currently calculated – this is another reason why the SEM presently used is an underestimate. For reasons related to protection of the public, the reviewers think it is appropriate to maintain the adjustment at 1.64 SEMs but using a more appropriate estimate of the SEM; implementing this change is independent of that.

4. ***Conduct regression analyses to identify stations/examiners with possible aberrant characteristics.***
Currently, station-total correlations for all circuits combined and for individual circuits are calculated for each day of test administration to identify stations that are performing aberrantly. Because correlations are affected by both the strengths of relationships and the variability in scores, low correlations are somewhat difficult to interpret. A better approach would be to use regression analyses for this purpose. The figure below provides an illustration for two stations. The Y-axis is the expected station deviation score – the expected difference between a candidate's score on the station and the standard for the station. The X-axis is the "corrected" total test deviation score – the mean of the station deviation scores omitting the station under study; this (roughly) places total scores from different circuits onto a common scale (even if they

occur on different test dates with different sets of stations) indicating the amount that performance is above or below the pass/fail standard. Each regression line[8] on the graph represents a different examiner marking the station. The graph on the left depicts a station marked fairly consistently across examiners, as indicated by the small vertical spread in the regression lines. The graph on the right depicts a station that examiners are marking somewhat differently, with a set of regression lines (hawk examiners) clustered together low on the Y-axis and a pair of regression lines (doves) somewhat higher. The pattern suggests that there may be problems with the marking criteria used for the station; in addition, candidates may be advantaged (or disadvantaged) by the examiner marking the station[9].



Some variation in examiner stringency and in the spread of scores will reflect issues in station design, case content, and marking criteria. It would be interesting to hear more about the procedures in current use, but we consider that the outlined procedure should be more effective because it is conducted within the context of individual stations and allows for pooling of information across dates of test administration.

[8] The regressions depicted in the figure were done in a manner that constrained estimated regression lines to be parallel. While this is not necessary, it may be desirable because of the instability of estimated regression slopes with small sample sizes.

[9] This is the major reason to consider adjusting scores to reflect differences in examiner stringency as discussed in the next recommendation.

5. ***Modify test administration procedures to increase the "connectedness" of the design and obtain better estimates of candidate proficiency, station difficulty, and examiner stringency.***

Because a separate group of 13 candidates rotates through a separate circuit with its own set of examiners, there are really three distinct "test forms" on each day of test administration, with the same stations used in each test form. Because of the relatively small number of stations per circuit, differences in test (circuit) difficulty will occur even with random assignment of examiners to circuits[10]. A modification in the test administration procedure can rectify the situation. Rather than structuring the administration so that candidates and examiners are assigned to circuits, replicate the stations in the same sequence and have examiners (paired with the same role player) rotate from room to room. This will result in all candidates seeing the same set of stations in a "connected" design that will permit more accurate (unconfounded) estimation of candidate ability, station difficulty, and examiner stringency. This would make it possible to adjust candidate scores for differences in examiner stringency, which should improve the reproducibility of scores and pass/fail decisions by adjusting for differences in the stringency of examiners marking the same station.[11]

Connectedness means that the pattern of overlap in a dataset allows for unconfounded estimation of effects. In a large-scale OSCE, for example, if there are separate circuits in which different groups of examiners assess different sets of examinees, the dataset is disconnected, meaning that there is a confounding of examiner stringency and examinee ability.

What we are suggesting is a simple approach to connecting their disconnected design. Each day of test administration would still be disconnected, but circuits within a test date would be connected. This would make it possible to estimate station difficulty separately from examiner stringency and place examinee proficiencies on a given date of test administration onto the same scale.

[10] This is one reason that use of coefficient alpha yields over-estimates of reliability. The fact that there are three separate circuits is "invisible" to calculation of coefficient alpha, and systematic differences in rater stringency from one circuit to another, as a result, are assigned to true-score variance in computations. See the "thought experiment" (on pages S18-19) in Swanson and van der Vleuten (2013) for further discussion.

[11] This approach to test administration can be implemented independent of test length and standard setting methodology as long as the total number of stations is a multiple of the test length.

RCGP proposed action:

Comment on recommendation

Currently the three simultaneous circuits run independently, with examiners and role players remaining on the same circuit throughout each exam. HPAC recommends that we break this system to 'de-nest' or 'connect' those three groups, thus creating a single exam form. That might increase the accuracy of estimates of the variables discussed earlier. Benefits would be particularly apparent if we adopt the use of G-theory.

Operationally, this would involve examiners moving between circuits, to varying alternatives, part way through exams. This implies significant operational change, with potential to be disruptive to the smooth running of the exam.

Planned action

To run a feasibility pilot and seek further psychometric advice

Plan to Review

After the feasibility pilot to assess whether the gains in improving accuracy of QA processes would be outweighed by operational costs and difficulties.

6. ***Explore the use of key-feature style checklist items in combination with global rating scales for station scoring to decrease variability in marker stringency and increase consistency in the marking criteria used.***
While research on global rating scales has produced encouraging results, it seems likely that examiners vary in both criteria used and in stringency when they are used in isolation[12]. This may not be a major problem in small-scale OSCEs where all examinees are marked by the same examiners, but it can be a problem in large-scale OSCEs involving multiple circuits with several examiners marking the same station, particularly if the standard setting procedure does not take marker stringency into account. The Australian Medical Council and the Medical Council of Canada have developed an approach to marking that involves small numbers of key-feature-style checklist items used in combination with global rating scales. RCGP may wish to contact them for additional information.

[12] See pages S20-21 in Swanson and van der Vleuten (2013) for further discussion.

7. ***Review the ethnic diversity of case palettes***
   In the interests of inclusivity and representation, it would be useful to ensure that in each palette of cases, the mix of patients better represents the UK population.

RCGP proposed action
Response to comment
In each day's 'palette' the CSA has cases where diversity is a key element. The percentage of Role Players (RP) from BME groups is greater than that in the UK population, and the exam has recently requested increased recruitment of RPs from relatively under-represented groups.
The CSA has previously tried, and abandoned as impractical and unsuccessful, efforts to run cases through translators and also to have actors with heavy accents. This HPAC recommendation, as explained by the authors, relates mainly to increasing the diversity of names e.g. adding more Central/Eastern European names, and also to reducing the association of names with an ethic group or with a disease pattern. As a result, patient names would be irrelevant to the content of the case.
Planned action
To progressively amend the names on the cases during the case review process, being mindful of the need for a palette to represent the entirely of the UK population.

8. ***Review the make-up of the examiners' panel***
   In the interests of inclusivity and representation, it would be useful to ensure that this is representative of the UK population of GPs particularly with respect to age.

RCGP proposed action
Response to comment
A process to select examiners from groups representative of the UK population of GPs has been in place since 2010. The composition of the panel was reviewed recently, and women, younger doctors and International Medical Graduates remain relatively underrepresented. A recent advertisement for recruits was successful in recruiting those with characteristics currently under represented on the panel. The minimum experience as a GP was reduced from five to three years WTE GP experience to attract younger applicants. Our primary responsibility is however to exam candidates, thus competence as an examiner is paramount. The necessary skill set overlaps significantly with that of educators. Educators as a group are not representative of the UK population of GPs.
Planned Action
To continue to aspire to the panel of examiners representing the make up of UK GPs as closely as possible.
Plan to Review
Review success of recruitment of under-represented groups regularly.

9. ***Explore incorporation of role-player ratings into marks for communication skills.***
   With some training, lay persons (trained role players) have a good perspective for marking communication skills; incorporating a public perspective seems a legitimate contemporary aspect of stakeholder involvement. This is widespread practice at undergraduate level.

RCGP proposed action

Response to comment

All marks awarded to candidates come from trained GP examiners. Role Players do not currently contribute to the assessment. There is encouragement from the GMC to increase lay involvement in assessment and other colleges are already doing so. Increasing the lay and/or patient voice was already an area of investigation prior to this review. An exploratory pilot during the review process investigated the performance of both lay assessors and role players, estimating their agreement with GP examiners and modelling the effect on candidate marks (hence outcomes).

Planned action

To continue to explore the feasibility of RP marking considering what feedback would be most helpful to candidates.

Summary

The reviewers were impressed with the extensiveness, clarity and transparency of the information available on the RCGP website and the willingness of the various individuals who made time to speak with us and share information relating to the examinations.

Overall, the reviewers thought that the CSA and AKT were fit for purpose and fair for both candidates and patients. The 'new format' examination, when introduced in 2007, was a 'state of the art' programme of assessment. We have made some recommendations motivated by the potential to enhance the validity and defensibility of the examination system, in line with developments in assessment theory and practice over the last ten years.

In relation to the Kane Validity Framework, our views are as follows:

The first stage of the Kane Validity Framework (KVF), i.e. the statement of purpose of the examinations, is clearly and explicitly stated, for the overall programme of the MRCGP assessments, as well as for each component.

The second stage of the KVF – the five domains of validity evidence: the reviewers were able to find all the evidence required for to evaluate each of the domains, and it was clear that the RCGP had considered the various aspects in relation to designing and implementing a programme of assessment which met international standards.

Stage 3, the Interpretive Argument (how all the components explained in the evidence section come together to form a strong case for using the assessment, from a validity perspective) is also adequately addressed.

Overall, the team of reviewers considered that the AKT and CSA components of the MRCGP met, or in some places, exceeded standards for procedures used for high-stakes examinations in the medical profession.

This view was unanimously agreed by all the reviewers in relation to both national and international perspectives, based on their extensive experience of current best practice and the medical education literature. We considered that MRCGP in its current form (and with suggested developments), is in line with the GMC's SCAR recommendations and Generic Professional Capabilities requirements.

References

Cook D.A., Brydges R., Ginsburg S., Hatala R. (2015) A contemporary approach to validity arguments: A practical guide to Kane's framework. *Medical Education*. 49(6):560-75.

Clauser B.E. et al (2017) Issues of validity and reliability for assessments in medical education. In E Holmboe et al (Eds) *Practical Guide to the Evaluation of Clinical Competence*, 2nd ed. Elsevier.

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

Guba, E.G. (1989) Lincoln YS. *Fourth generation evaluation*. Newbury Park, CA: Sage.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5-17.

Kane, M. T. (1994). Validating interpretive arguments for licensure and certification examinations. *Evaluation & the Health Professions, 17*(2), 133-159.

Kane, M. T., Crooks, T. J., & Cohen, A. S. (1999). Designing and Evaluating Standard-Setting Procedures for Licensure and Certification Tests. *Advances in health sciences education: theory and practice, 4*(3), 195-207.

Kane M.T. Validation. In: Brennan RL (Ed) *Educational Measurement*, 4th ed, Westport, CT: Praeger 2006;17–64.

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement, 50*(1), 1-73.

Kemp, S.J. (2012) Constructivist criteria for organising and designing educational research: How might an educational research inquiry be judged from a constructivist perspective? *Constructivist Foundations.*8(1):118-125.

Newton, P. E. (2013). Two Kinds of Argument? *Journal of Educational Measurement, 50*(1), 105-109.

Norcini J.J. et al (2011). Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. *Med Teacher* 33:206-214.

Sireci, S. G. (2013). Agreeing on Validity Arguments. *Journal of Educational Measurement, 50*(1), 99-104.

Swanson, D.B. & van der Vleuten, C.P.M. (2013) Assessment of clinical skills with standardized patients: State of the art revisited *Teaching and Learning in Medicine,25: sup1, S17-S25*

van der Vleuten C.P.M., Schuwirth L.W.T. (2005) Assessing professional competence: From methods to programmes. *Medical Education* 39:309–317.

This Report was prepared by Katharine Boursicot BSc MBBS MRCOG MAHPE NTF SFHEA FRSM

Director
Health Professional Assessment Consultancy

Appendix 1: Summary of the key features of the Kane Validity Framework

The latest iteration of the Kane model of validity has three stages (Michael T. Kane, 2013; Newton, 2013; Sireci, 2013):

1. Statement about the **intended purpose/use** of the assessment
2. Gathering **meaningful evidence** to support the interpretation of the test results
3. Rationalising and interpreting the evidence into a convincing **argument** to justify the decisions made (Newton, 2013)

Stage 1: Statement of Purpose

This section relates to the purpose of this assessment, what domain(s) of competence are being tested and how it connects with other assessments in the course/programme.

Stage 2: Evidence

The model identifies five key domains of validity evidence (Downing, 2003; M. Kane, T. Crooks, & A. Cohen, 1999; M.T. Kane, 1994; Michael T. Kane, 2013; M. T. Kane, T. J. Crooks, & A. S. Cohen, 1999). For each validity domain any assessment strategy should specify minimum requirements, as follows:

1. **Assessment content:** The rationale for choice of test formats and their suitability for the learning outcomes to be assessed (e.g. MCQ, OSCE, SAQ, WPBA, EPAs, etc.); formal blueprinting of each test's content to learning outcomes to achieve balanced sampling; sufficient sampling; good test item design; internal review of test items at pre-test and post-test.

2. **Assessment response process:** Ensuring candidate familiarity with test formats; examiner training in scoring/judgement methodology; quality control of the collection and processing of scores and judgments.

3. **Internal structure of assessment:** Post-test analysis of whole tests and of individual items. Importantly, this includes test reliability (reproducibility of scores) and standard error of measurement (to indicate confidence intervals to pass-fail cut scores); formal standard setting and the application of pass-fail rules to scores and judgements; clear interpretation and reporting of scores and judgements for candidates.

4. **Relationships to other variables:** Consideration of the relationships between performances of the same candidates on different tests. This may involve correlations between tests taken earlier and later in the in educational development (predictive validity) and between tests taken at about the same time (concurrent validity).

5. **Consequences of the assessment outcome:** Consequences for candidates, society at large and for the awarding institution of pass-fail outcomes; reasonableness and reliability of pass-fail determination; equity and fairness; extenuating circumstances and appeal procedures.

Stage 3: Interpretive Argument

This section relates to how all the components explained in the preceding sections come together to form a strong case for using the assessment, from a validity perspective.

The Kane model of validity and the specified evidence requirements should inform the basis of planning and developing any framework of assessment and be supported by clearly written regulations, the provision of appropriate information to candidates and by training, including refresher training, for examiners. This model of validity also provides a framework for quality assurance purposes, as it draws together many elements relating to examinations in a structured and coherent manner.