MRCGP AKT

Fairness Review (FR) September 2025

Summary of Action Points

AKT Exam content

- **Fairness**: Review whether certain exam questions are fair to all candidate groups, particularly those where differential item performance (DIF) was observed.

  Action: ongoing work and all questions showing statistically significant differential item performance are reviewed after each exam. Those showing a high DIF, or which cannot be more easily explained*, are flagged for the next annual Fairness Review (scheduled for October 2026).

- **Question Complexity:** Review the phrasing and complexity of exam questions, particularly those involving statistical terms and relative increases, to assess if language or calculation requirements unfairly disadvantage international graduates or those with different language backgrounds.
  Review the impact of complex language in AKT questions on candidates whose first language is not English and consider potential adjustments to improve equity without reducing exam rigor.
  Consider rephrasing terms in exam questions such as "best prediction" (and other terms identified in the groups) to simpler, more familiar terms for candidates.

  Action: Expert linguist review in 2025 (summary can be found here) has reviewed the language use and we use this to inform our item writing processes. Careful attention is given to sentence construct and length, the avoidance of superfluous wording and reading ease. There are intentionally few items with long scenarios. It should be noted that a previous external independent linguist review also demonstrated reassuringly low IELTS score for AKT items. We will specifically look at statistical terminology in our upcoming item writing.

- **Geographical Variation in Exam Performance:** Investigate whether candidates from different geographical areas perform variably on questions related to local cultural knowledge, such as benefits eligibility, and analyse if this correlates with exposure in practice.

  Action: consider this as a possible future research question.

Dr Roisin Cartwright-Terry FRCGP
AKT Clinical Lead

- **Presentation of Medication Options in Exam Questions:** Evaluate alternative ways to present medication options (e.g., using abbreviations, bullet points, or grouping) in exam questions to reduce cognitive load, especially for candidates with specific learning difficulties or reasonable adjustments.

  Action: increasingly, medication lists are provided rather than in sentence format. It is often important to include full names not just abbreviations however to avoid disadvantaging some candidates. We will continue to be mindful of when this is helpful or not.

- **Highlighting Abnormal Results:** Evaluate the current practice of identifying abnormal lab results in AKT questions and determine if changes are needed to better reflect real-life clinical decision-making and avoid missed diagnoses.

  Action: review into this has taken place. Abnormal results are not highlighted in the AKT for a variety of reasons including limitations and consequences to candidates of marking very borderline results as abnormal. We now do give normal ranges where possible next to results. As a result of the FR25 feedback we have raised the discussion again but at present continue as previous.

*Most highlighted item DIF is relatively easily explained, for example:

• Women tend to score higher marks than men on female health questions such as contraception, and on infant heath questions. This is likely due to the types of consultations women see more frequently in primary care, and their own personal family experiences

• UKGs tend to score higher marks on data interpretation questions. This is likely due to differences in undergraduate and postgraduate training experiences, and, despite clear GP Curriculum statements, a feeling of uncertainty for some IMGs why such questions are included in the exam

• IMGs tend to score lower marks on organisation and management questions. This is likely due to having had less experience of UK primary care culture and practice.

Outline of the day

Background

An RCGP advert asking for volunteers (expenses remunerated) was sent by email cascade to stakeholders and educational provider leads. Fifteen GP Registrars and Newly Qualified GPs, who had passed the AKT, were randomly selected based on a spread of demographic and geographic factors to ensure a mixed group.

This was a routinely scheduled, annual quality assurance review of the fairness of content and question style within the MRCGP AKT assessment.

Process

1. Participant confidentiality and consent agreements

2. Background and rationale for annual Fairness Reviews

3. Review of FR 2024 Action Points - all have been actioned or progressed, except for the publication on the AKT website of an example scenario with an explanation of specific lead-ins. This is now completed.

4. Independent psychometric expert explanation of differential item analysis (DIF) and the process of individual item review

5. Question and answer session

6. Two separate small group workshops to each review a different set of 15 paired questions. Both groups were initially blinded to which question of the pairing showed significant DIF and which did not. The DIF questions were identified by the independent psychometric team from the previous 12 months of AKT exams.

7. The two groups were:

    a. led by facilitators

    b. recorded with consent

    c. observed by three AKT Core Group observers

    d. observed by the independent Psychometric expert

8. Themes and priorities were collated by the two small group facilitators and checked for accuracy with group members

9. Whole group summary of views, discussion and action points agreed

Observers Report

Nomenclature

IMGs	= International medical graduates
UKGs	= UK graduates
RCGP	= Royal College of General Practitioners
MRCGP	= Membership of the Royal College of General Practitioners
AKT	= Applied knowledge test
ANP	= Advanced nurse practitioner
CCT	= Certificate of completion of training
SpLD	= Specific learning disability (eg dyslexia)
ST 1/2/3	= Specialist training (year 1/2/3)

Group members

15 representing a wide range of ages, backgrounds, and geographical area of practice.

All had passed the AKT within the last five years.

Were mostly ST3 but some were post-CCT or ST2

GROUP DISCUSSION AROUND THE SLIDE PRESENTATIONS OF DIFFERENTIAL ATTAINMENT:

WHAT WERE THE COMMON THEMES?

Exam rigour

The group members felt that the AKT should reassure patients on safety and that it should be sufficiently rigorous to engender respect from specialist colleagues.

How experience and exposure affect exam performance

The group members suggested that familiarity with conditions affects exam performance, and that exposure to clinical situations drives learning.

Familiarity with medical problems and administration may come from a variety of sources including personal and family experience, as well as clinical exposure and teaching. In the questions we reviewed, this sometime favoured UKGs, sometimes IMGs, sometimes female candidates, and sometimes male.

Discussions included the greater exposure of IMGs to personally managing medical problems such as diabetes and infections. The groups felt that IMGs might have more hands-on experience and were more comfortable with interpreting physical examination findings.

We were also told that they were more likely to have broader clinical experience, including other medical specialties. They might, therefore, recognise secondary care problems and rare diseases better. They may also be more used to taking prescribing decisions outside of established guidance. On the other hand, UKGs benefit from community-based teaching on common primary care problems. Occupational disease may be less well recognised in countries with less industrial regulation. Mental health problems and dementia may carry more stigma, and exposure to their diagnosis and management will be less. This might be compounded by different diagnostic approaches and the use of different terminology. UKGs have not only been taught in the NHS system but have often also had life-long experience of the service. They may also have direct or indirect experience of the welfare system. IMGs were thought more likely, however, to undertake mandatory training and to selectively revise administrative topics, helping to fill some of the knowledge gaps.

In previous fairness reviews our participants have told us of the increasing challenge posed when activities are routinely delegated to other team members. In this review we were told that chronic disease care is often managed by practice nurses or, in the case of diabetes, GPs with an interest. We were also told that male trainees are called upon only infrequently to manage women’s health problems.

Questions testing evidence-based medicine and national guidance might suit IMGs, as they were likely to consult relevant sources more frequently to compensate for gaps in their training. We were told that local guidance may differ and access to resources or referral pathways may influence the roles and responsibilities of doctors in different regions. It was felt that UKGs may be more influenced by this.

One group told us that cultural differences relating to lifestyle may adversely affect how questions are answered. Examples we were given include culturally specific approaches to infant feeding and weaning, meat eating and the types of vegetables consumed. If these are not critical to the question, then it was suggested that alternatives are found. They felt that units of alcohol were not always easy to understand when candidates had little or no personal experience to draw on.

How language complexity affects performance

Everyone who attended had been working in the UK for at least a few years. There were different views on whether the language used detracts from the knowledge assessment in the AKT. They highlighted that some items that may cause difficulty in exams for candidates whose first language is not English.

The group members supported the use of simple language. They stated that the AKT is an applied knowledge test rather than an applied language test. They felt complexity led to differential performance. They suggested that candidates whose first language is not English require additional cognitive steps to analyse and process questions. This leads to difficulty in completing the AKT in the allotted time (this might have eased a little with the recent introduction of a shorter exam). IMGs often find medical jargon less ambiguous than colloquialisms as that reflects how they have been taught. This is an example of where attempts by the AKT to simplify language may backfire. Some common medical terms may also be unfamiliar to IMGs e.g. ‘cognitive impairment’. They highlighted several statistical terms and phrases that could be challenging: ‘detection rate’; ‘positivity rate’, ‘best prediction’ (though this isn’t a common AKT phrase); ‘which number best represents the risk’; ‘relative increase in admissions expressed as a percentage’. Negatively phrased options also increase complexity. These are rare in the AKT but more common in question banks.

The group members were keen to point out that long question scenarios are not necessarily associated with complex language, though both may affect reading times for IMGs and those with SpLDs. They stated, however, that simplified language had to be balanced against producing an authentic exam and one that best ensures safe practice.

How the style of questions affects performance

Factual questions are often simpler, easier to read and quicker to answer. The group members thought that IMGs may find these questions easier as this matches the style in which many have been taught. In contrast, it was felt that UKGs are trained more in interpretation of situations than in pure factual recall. Postgraduate training also encourages experiential learning and reflection. The group members felt that factual questions should be valid and relevant to daily practice. Doctors may look up

an answer when they deal with a less common problem. While questions on rarer problems test the breadth of the GP curriculum, they may not always be important to instantly recognise.

The group members discussed more complex question formats. Two stage questions might present symptoms but ask for the correct management option, which requires an intermediate step of making a diagnosis. Multifactorial questions may include a scenario and then add a graph or a table. They noted that these formats are visually complex and are harder to comprehend. They thought they take longer to answer and may disadvantage IMGs and those with an SpLD. However, they also felt that such questions might reflect a real-world approach.

Group members felt that it was helpful to add the phrase “according to national guidance” at the beginning of the question. They considered that it prompted them to recollect specific guidelines. They told us that IMGs are more likely to use books and guidance to revise their knowledge. They also suggested that IMGs are less influenced by local protocols and specialist opinion than UKGs especially when it differs from NICE/SIGN.

Although we mostly concentrated on language, the group members also commented on charts and graphs that used multiple colours. They pointed out that this wasn’t always helpful. As a footnote, the AKT does actively check that colour combinations do not disadvantage those who suffer from forms of colour blindness.

Candidates with dyslexia (the commonest SpLD) said they find numbers and graphs easier to interpret. Unsurprisingly, data interpretation questions with long descriptions were felt to be harder. Where the written introduction was longer, those with an SpLD could still answer correctly but it would take more time. Some thought that IMGs might overthink these questions, wanting to read carefully all the information rather than focusing on the graph. Some of the IMG group members discussed how they processed the information: they would read the text, interpret (into their thinking language), look at the graph and then decide on how to approach the question. These steps all take extra time. Some also suggested that men, related to a different approach to risk, might perform better on these questions than women. The AKT have agonised over whether charts and graphs should be true to the original. On the whole, the group members felt that accuracy was less important than reduced visual complexity.

Repetition in questions can overwhelm those with an SpLD, increasing the time taken to answer. Repeated numbers and mixed number and word formats were particular concerns. How we present medications may also cause processing difficulties. Long lists of medications in single sentences and use of the full names rather than common acronyms (e.g. LABA) both increase cognitive load. The AKT will usually provide both the full name and the acronym.

There was debate on how the results of investigations are presented in tables. Primary Care IT systems often highlight abnormalities, even if they are not clinically relevant. Equally, normal tests can sometimes be important. The group members felt that if abnormalities are regularly highlighted then, in an exam situation, important findings may be overlooked. They also debated if highlighting abnormalities was more useful than providing reference ranges, which is the current AKT policy. Some investigations are not suitable for highlighting e.g. microbiological testing, and reference ranges vary both around the UK and around the world. On the whole the group members supported the current AKT policy.

Providing good quality pictures can be difficult. One question presented a problem more common in BAME people, but the picture was of someone with white skin.

IMG group members felt that knowledge of the condition and an accurate representation was more important than the source of the picture. Whilst the AKT will strive to find the best and most appropriate pictures, it was reassuring to know that the picture didn't appear to disadvantage candidates.

We came across a similar concept in a question about a condition familiar to both UKGs and IMGs but with different aetiologies. The IMG group members didn't feel this affected performance as the presentation of the index problem was the same.

'Doing nothing' appears to be a difficult clinical concept for IMGs in an exam setting. We were told that non-UK exams generally don't test for the ability to identify normality.

To avoid confusion, group members felt that questions should clearly specify if the goal is to improve prognosis or symptoms where this can affect the correct answer.

Approaches to education and training and how they might change

There was consensus that training schemes should help trainees identify and address gaps in knowledge and experience. Although there appear to be specific themes, learning needs and learning styles are individual.

Group members felt that the greatest educational benefit from trainers and training schemes was to prepare candidates for more complex questions by providing plenty of experiential and qualitative learning, and to identify and plug gaps in learning.

Factual questions may lend themselves to rote learning; our group members also said that question banks help to prepare candidates for these questions. They told us that some question banks are targeted towards IMGs, often focusing more on prescribing.

Group members suggested that IMGs and UKGs have different strategies: UKGs may rely on experiential learning and targeted revision whilst IMGs use a wider range of resources, focusing on high yield topics. On the subject of targeted learning, the group members felt that time-efficient learning could focus on the areas with a 'higher return' e.g. management and treatment and using the AKT feedback reports to identify knowledge gaps. We were told that IMGs are more likely to join dedicated group study sessions either locally or on internet platforms e.g.14Fish. Local study sessions were organised by themselves or facilitated by experienced GPs and trainers. Specific groups were often organised for trainees who had previously failed exams. Study group sessions covered specific topics, shared strategies, and addressed areas of weakness. Some Deaneries may provide exam preparation packages or access to resources, but the amount of help given varies around the UK. The group members also suggested that exam experience helps candidates spot 'red herrings' (though the AKT would not intentionally use such a strategy in question development).

The group members discussed that statistics and data interpretation teaching was a feature of UK university curricula (including UK universities abroad), although the emphasis may differ from one medical school to another. They suggested that those who are part of journal clubs or had intercalated degrees may acquire greater skills. Access to practice question banks on relevant statistics would help IMGs who have missed out on acquiring these skills in undergraduate training. They felt there should be a greater emphasis on statistics and interpretation in training. In particular, some concepts e.g. Odds ratios are likely to be more unfamiliar to IMGs.

They told us that formal training sometimes overlooks gaps e.g. GP interactions with the processes of the welfare state. A few organisational questions are also relevant to those who hope to fulfil leadership roles e.g. partnership. Our group members told us that experience and training in these areas varies. Sometimes, mandatory training may compensate, but they thought that trainers and training schemes ought to be mindful of this when considering how to cover the breadth of knowledge needed for the exam.

Differential attainment that we were unable to explain

We presented a pair of questions with differential attainment by candidates with 'Reasonable adjustments' (but not SpLD) i.e. the awarding of extra time for a range of other disabilities. Our group members were unable to explain why this difference had occurred. In previous fairness reviews we have also been unable to explain this phenomenon. As the position of a question in the test is randomised it was unlikely to be due to failing to finish in time.

Other discussions: exam content

Some questioned whether the topic of welfare benefits was relevant to a GP exam especially as it wasn't universally covered during training.

Overview of group discussions on how differential attainment should influence AKT development

- Participants debated the challenge of determining acceptable degrees of differential attainment, emphasizing the need for careful question design to avoid overcompensation and ensure validity, with recognition that some differences may be inevitable but should be contextually justified.
- Participants warned that focusing on a single aspect of a question to address differential attainment may result in unintended bias in the opposite direction, emphasizing the need for holistic review and careful standard setting.
- Acceptable Degree of Difference: The group acknowledged the difficulty in determining what constitutes an acceptable level of differential attainment, noting that validity must be contextually assessed and may vary depending on the question and candidate group.

Prioritisation exercise

A list of important themes identified by the facilitators from the morning groups were presented to all attendees in a combined session in the afternoon. Group members were asked to score and rank these themes according to importance in terms of changes to the exam. After this exercise the group were asked to justify their scores. They were then asked to rank their scores again.

<u>Round 1</u>

The two **most important themes** were adjudged to be:

'linking questions to national guidance'

'clear layout of questions'

The **two least important ranked themes** were adjudged to be:

'avoiding questions where lack of cultural exposure may disadvantage'

'avoiding use of repetition in questions'

<u>Round 2</u>

The **highest ranked score** was now:

'ensure a wide coverage of the curriculum'

Items that **significantly decreased** in score after the discussion were:

‘avoiding unnecessary complex language’

‘factually based questions’ decreased’

The peer discussion had a noticeable impact on the consensus for priorities for the exam.