Likelihood of examination success relating to exam re-sits, timing and frequency and patient safety implications of multiple re-sits

Review of evidence

John C. McLachlan
Jcmclachlan1@uclan.ac.uk

Index

Executive Summary

Headlines

The current RCGP process allowing multiple re-sits within a short time scale, followed by a ban on further attempts is not optimal in terms of identifying those who ought to pass or fail. It would be better to allow a maximum of four attempts, followed by a mandatory re-training period before allowing further attempts.

1 Key Concepts

Assessments are designed to explore an underlying latent trait or construct in candidates. They should be valid (measure the thing you want to measure) and reliable (measure consistently). 'Fairness' is often confused with unreliability, but may mean either 'people should have equal outcomes', 'people should get what they deserve', or 'people should be compensated for bad luck', and these can be at odds with each other. There is no sharp dividing line separating 'true passes' from 'true fails'.

2 Do high stakes professional exams predict later clinical practice?

There is clear evidence from a variety of sources that performance on national licensing exams is a statistically significant moderate predictor of performance in later clinical practice, by a variety of measures and outcomes. However, there is a great deal of unexplained variance (perhaps as much as 80% to 90%), and the predictions are not specific at the individual level.

3 Why do trainee GPs fail their exams?

Reasons why trainees may fail assessments such as the AKT and CSA are complex and may include construct irrelevant factors. From the candidates' perspectives, these may include the interaction between country of training and ethnicity, gender, age, personality characteristics, financial circumstances, and socio-economic background. The assessment itself may be unreliable, and assessors may 'fail to fail' in advising candidates on readiness to sit the exams. Guidance might be offered to supervisors on determining when trainees are truly ready to sit, and a 'sign-off' of readiness to undertake the assessments introduced.

4 What is the relationship between re-sit performance, number of re-sits and the properties desired of those succeeding?

Scores generally improve on re-sit, by as much as 0.3 or 0.4 Standard Deviations, but the improvement decreases with each attempt, and may plateau after two or three attempts. Scores can improve on re-sitting due to three factors: familiarity with test material, statistical variation, and improvement on the construct under test. Short tests of ability show most effect of familiarity; improvement on longer credentialing or achievement tests

suggests that there is improvement on the test construct. Candidates on longer achievement tests do not receive an advantage from sitting the same test items again. Candidates improve on the test construct after further study.

5 Risks associated with passing poor performers

Both false positives and false negatives in professional medical exams have costs, both to the individual and society. False positives pose risks to patient safety, to the extent that safety errors are caused by individual errors rather than system factors. False negatives deprive society of needed doctors in times of shortage (and in the context of this report, deprive the NHS and public of GPs).
A limit is recommended of four attempts followed by additional training. Without re-training, passes after more attempts are likely to be false positives. However, improvement on scores following further training, is likely to represent a true improvement.

6 Medical Error and Safety

Medical error and patient safety are not simple matters of individual errors by doctors, which can be addressed just by raising barriers in medical professional exams. In particular, allowing fewer doctors to qualify for a profession or speciality raises the risk of under-staffing, which is in itself a major source of medical error.

7 Re-training

Receipt of structured feedback and targeted training during a mandated 'refractory' period is essential to improvement on the property under test.

8 Conclusions and Recommendations

Current evidence shows that the number of permitted re-sits should be limited, since those who have not passed after three or four attempts are not likely to pass on an immediately subsequent occasion. Perversely, permitting further attempts after this number is likely to pass only True Negatives, who pass through chance, test familiarity and the variability of the exam. However, there is clear evidence that performance can improve on further training and feedback. The optimal strategy would therefore be limit the number of re-sits to no more than four, then to require a mandatory period of targeted training of at least one year, and more plausibly two, after which a further limited number of re-sits should be permitted. By the same logic, the number of further attempts should also be limited, to a maximum of two, since candidates should be able to demonstrate that significant improvement has taken place since the last exam in a maximum of two attempts.

Concerns may be expressed that this may allow doctors to pass who are not safe to do so. However, performance improvement subsequent to further training in an extended credentialing exam is evidence of an improvement in the construct under test. In any case, the numbers involved are likely to be small compared to the False Positives who passed the initial 'first sit' assessment.

Attention should be given to the preparedness of candidates to undertake the assessment in the first instance.

Report Author

The author is currently Professor of Medical Education at UCLan and was previously Associate Dean of Medicine at Durham, and Foundation Director of Phase 1 at Peninsula Medical School. He served as Editor-in-Chief of the journal *Medical Education*, and was a Raine Distinguished Visiting Professor at the University of Western Australia. He has undertaken research projects for the GMC, HCPC, Department of Health, and Scottish Government, on high stakes medical assessment in general, and the measurement of professionalism in particular. The GMC amended its policy on re-sits for PLAB in part on the basis of his input. This particular report for HEE draws on his previous research as well as new research undertaken for the project. Although HEE were offered the chance to comment on the draft report, and to identify any errors or misunderstandings, it is an independent piece of work, and the author has no conflicts of interest to declare.

1 Key Concepts

It is essential to first define some key concepts and technical terms, particularly since some terms have technical meanings which are different from their 'everyday' language sense.

Validity

Validity means 'does the test measure what you want it to measure?'. In the context of the RCGP exams, the *latent trait* (see below) under study is that of fitness to work as a GP.

Validity comes in a variety of different forms. For the purposes of this Report, the key forms are probably *face validity* ('do the individual components of the test look appropriate to the relevant experts and stakeholders?'), *content validity* ('does the test as a whole cover all the areas that it ought to cover?') and *predictive validity* ('Do the tests predict how candidates will perform in later clinical practice?').

Reliability

Reliability in this context merely means 'does the test measure with consistency?', as opposed to common meanings of 'reliable', or 'able to be relied on'. A test can therefore be reliable but not valid, if it is measuring the wrong thing. Again for the purposes of this report, three kinds of reliability will be briefly mentioned. *Classical test theory* suggests that an observed score is made up of a true score and an error score. Typical classical test theory measures are Cronbach's alpha and the Standard Error of Measurement. *Generalisability theory* operates on the basis that error can be divided into various factors such as candidate variance, assessor variance and item variance, and the interactions between them. It is generally measured by the co-efficient G. *Item Response Theory* is based on the relationship between the latent trait and performance on the test, and is often used in high stakes, large scale, testing environment.

Fairness

While it is often stated that tests should be 'fair', it is less often indicated exactly what is meant by this. One common usage actually relates to *reliability*: candidates complain that a test 'is not fair' if there is evident variability between assessors or test occasions. However, 'fairness' can also refer to aspects of justice. In very simple terms, *strict egalitarianism* means that everyone should get the same. *Desert egalitarianism* means that everyone should get what they deserve. *Luck egalitarianism* suggests that people should be compensated for ill fortune that is not their fault. Unfortunately, these can be at odds with one another. Consider, for instance, gender ratios in medicine. Strict egalitarianism might be taken to mean that the ratio of males to females entering medicine should be the same as that of the population. But since females generally outperform males at school, those might require discrimination against better qualified females in favour of less well qualified males, during selection for medical school. Is this therefore at odds with desert egalitarianism? Or, under luck egalitarianism, shouldn't males be compensated for their ill-luck in being born male? Yet there is some evidence that female doctors are less likely to be sanctioned by the

GMC[i], or referred to NCAS[ii], may have better mortality outcomes than males[iii], and are generally paid less[iv]. Would it therefore be appropriate to say that if they were admitted to medical school in higher proportions than males, that this would be their just deserts? I have chosen the example of gender to illustrate these challenges, although I could also have chosen age: as will become clear, there are issues of ethnicity and country of primary medical training arising during GP training and assessment which raise similar challenging issues.

Latent traits and constructs

It is possible to imagine that examinee performance on a test can be explained or predicted by the presence of a defining characteristic of the examinee. Since these characteristics are not directly observable, they are often referred to as *latent traits* (Hambleton and Cook, 1977)[v].

Similarly, the aspect which is desired to be tested may be known as the *construct*. There is an extensive and complex literature on construct validity, which I will not enter into here, preferring to focus on face, content and predictive validity. However, it is important to distinguish between *construct relevant* and *construct irrelevant* factors. Construct relevant factors are those related to the latent trait under assessment, such as knowledge, skills (psychomotor and communication), and attitudes (the affective domain, including responding to feedback). Construct irrelevant factors are those not directly related to the latent trait, but which none the less impact on performance (generally negatively). For instance, a trainee who is stressed about financial difficulties, or who is a single parent with child care responsibility, may not perform as well as a trainee without these problems, but this does not mean the latent trait is less well developed. As will be seen, the degree of familiarity with a testing process can contribute to construct irrelevance.

Self-evidently, a candidate may fail an assessment because the latent trait is not sufficiently developed. So in a test of knowledge or skills, the candidate may lack the required knowledge, or be unable to perform the required skills. Of course, this may be entirely the candidate's responsibility – they may not possess the cognitive ability or physical or communication skills to perform well. But there are also issues about the assessment/training process – how reliable is the assessment, how effective is the training, how are candidates selected and progressed in programme? – which contribute to the measurement outcome.

Pass mark

This may seem a straightforward and familiar term. Yet it carries the implication that those above the pass mark deservedly pass and those below it deservedly fail. It is not so simple, as we will see below, and as a result, the term *cut score* is more accurate.

Pass/fail standards and standard setting

How is the cut score best determined? Case and Swanson (1996) are credited with saying "Standard setting is always arbitrary but should never be capricious". By this they mean

that the standard is set by arbitration between a group of experts, and is therefore always a social construct. There are no absolute standards that can be set prospectively.

It may be helpful to consider assessment as analogous to a screening test in medicine, and define terms accordingly:

True Positive (TP): Candidates who pass and deserve to pass
False Positive (FP): Candidates who pass and deserve to fail
True Negative (TN): Candidates who fail and deserve to fail
False Negative (FN): Candidates who fail and deserve to pass

We can then further define operational terms:

Sensitivity = TP/(TP+FN)
Specificity = TN/(TN+FP)
Accuracy = (TP+TN)/Total

No cut score divides candidates cleanly into True Positives and True Negatives, since the categories overlap due to (a) variation in candidate performance (b) variation in test performance from occasion to occasion, among other factors. It is possible to set the bar so high that all True Negatives will fail, but then so will a number of True Positives. Conversely, the bar can be set so low that all True Positives pass, but then so will a number of True Negatives.

Candidates want tests to be highly Sensitive (minimising False Negatives). But if False Positives are more expensive than False Negatives, then society may want a test to be highly Specific (minimising False Positives)

Since there is no ideal cut score that will eliminate all True Negatives, without also causing the loss of a significant number of True Positives, in practice, therefore, there is no cost-free solution that removes the risk of passing any candidates who will later pose a risk. There may be a need for a cost-benefit analysis. At what point is the potential cost of having a weaker GP exceeded by the potential cost of not having one at all? I will return to this in Section 5.

Summary: Assessments are designed to explore an underlying latent trait or construct in candidates. They should be valid (measure the thing you want to measure) and reliable (measure consistently). 'Fairness' is often confused with unreliability, but may mean either 'people should have equal outcomes', 'people should get what they deserve', or 'people should be compensated for bad luck', and these can be at odds with each other. There is no sharp dividing line that separates 'true passes' from 'true fails'.

2. Do high stakes professional exams predict later clinical performance?

This section draws on previous reports for the GMC and others. If professional exams such as those administered by the Royal Colleges do not predict later clinical performance (in other words, if they lack *predictive validity*) then their use as gate-keepers for entry to

specialities would be inappropriate. There are a number of good studies from various contexts which relate to this issue, including an excellent meta-analysis featuring data from the US Medical Licencing Examinations (USMLE, formerly National Board of Medical Examiners, NBME). In this review, Hamdy et al (2006)[vi] conclude:

"The studies included in the review and meta-analysis provided statistically significant mild to moderate correlations between medical school assessment measurements and performance in internship and residency. Basic science grades and clinical grades can predict residency performance".

The authors also concluded that, as might be hoped, performance on similar measurement instruments is better correlated than performance of different instruments. So NBME II scores correlate well with NBME III scores, medical school clerkship grades correlate well with supervisor rating of residents; and OSCE scores correlate well with supervisor rating of residents, when similar constructs are assessed. The results of their meta-analyses are extracted and summarised in Box A.

Box A (NBME is the previous version of USMLE)

Predictor	**Outcome**	**Correlation**	**Confidence Interval**	**Descriptors**
NBME I	supervisor rating during residency	Pearson r = 0.22	0.13-0.30	positive significant low
NBME II	supervisor rating during residency	summary correlation coefficient r = 0.27	CI 0.16-0.38	positive significant low
Clerkship Grade Point Average	supervisor rating during residency	Pearson r = 0.28	CI 0.22-0.35	positive significant low
OSCE	supervisor rating during residency	Pearson r = 0.37	CI 0.22-0.50	positive significant low
Clerkship Grade Point Average	supervisor rating during residency	Pearson r = 0.28	0.22-0.35	positive significant low
NBME I	American Board of Medical Speciality Examination	Pearson r = 0.58	0.54 – 0.62	positive significant moderate
NBME II	American Board of Medical Speciality Examination	Pearson r = 0.61	CI 0.51-0.70	positive significant moderate

Tamblyn et al (2002)[vii] compared the performance of 912 family physicians in Canadian licensing examinations with subsequent performance measured by a number of indices, such as appropriate prescribing, delivering continuity of care, and screening patients for serious illness. For instance, they noted that higher scores on drug knowledge were associated with lower rates of contraindicated prescribing (relative risk 0.88). They concluded "Scores achieved on certification examinations and licensure examinations taken at the end of medical school show a sustained relationship, over four to seven years, with indices of preventive care and acute and chronic disease management in primary care practice".

Tamblyn et al (2007)[viii] compared performance on the Canadian Clinical Skills Examination (CSE), which is similar to USMLE Step 2 CS. Candidates who lay two standard deviations below the mean for communication skills in the CSE were significantly more likely to be the subject of non-trivial complaint in later practice.

Papadakis et al (2008)[ix] found that low trainer evaluations predicted the likelihood of later disciplinary action by State Boards, and perhaps less intuitively, so did low scores on knowledge assessments. This is surprising because disciplinary action usually results from more complex issues than mere lack of knowledge.

Holmboe et al (2008)[x] explored the relationship between physicians' scores on the American Board of Internal Medicine's Maintenance of Certification examination and a variety of indices such as delivery of diabetes care, mammography and cardiovascular care. Their conclusions, like those of Tamblyn et al (2002), were stated unequivocally: "Our findings suggest that physician cognitive skills, as measured by a maintenance of certification examination, are associated with higher rates of processes of care for Medicare patients".

In a study in Ontario (Wenghofer et al, 2009)[xi], two hundred and eight doctors who took the Medical Council of Canada Qualifying Examinations Part I (Medical knowledge and clinical decision making) and Part II (Clinical Skills OSCE) and subsequently entered practice, were selected for study. Their clinical practice was assessed by peer examiners using a structured chart review and interview. Doctors in the bottom quartile of both Part I and Part II had at least a 3 fold increase in the likelihood of being rated as providing an unacceptable quality of care.

A recent study (Norcini et al, 2014)[xii] has come closest to the most desirable outcome measure – clinical outcomes for patients. These authors conclude that "After adjustment for severity of illness, physician characteristics, and hospital characteristics, performance on Step 2 CK [Clinical Knowledge Test of USMLE] had a statistically significant inverse relationship with mortality. Each additional point on the examination was associated with a 0.2% (95% CI: 0.1%–0.4%) decrease in mortality. The size of the effect is noteworthy, with each standard deviation (roughly 20 points) equivalent to a 4% change in mortality risk".

In papers with colleagues, I have demonstrated that performance on the GMC's Professional and Linguistic Board tests (both Part 1, written, and Part 2, OSCE) predict later clinical performance. The same is also true of numbers of re-sits required to pass, which also

predict the likelihood of referral to, and sanction by, disciplinary regulators (Tiffin et al, 2014; Tiffin et al, 2017)[xiii],[xiv] I will return to this in the discussion of re-sits.

It is therefore possible to conclude that there is a positive relationship between performance in assessments and later clinical performance. (This conclusion is slightly stronger than that reached by Archer et al 2016[xv]). However, it is important to note that the correlations reported in such studies are generally low, and there is much unexplained variance. For instance, a correlation of 0.3 would explain less than 10% of the variance affecting performance. Moreover, in such studies, those in subsequent difficulty may well be found to have performed poorly in prior educational settings, but that does not mean that all those who performed poorly in the educational setting are likely to perform poorly in later clinical practice. In terms of our earlier discussion, the test may be sensitive rather than specific.

Summary: There is clear evidence from a variety of sources that performance on national licensing exams is a statistically significant moderate predictor of performance in later clinical practice, by a variety of measures and outcomes. However, there is a great deal of unexplained variance (perhaps as much as 80% to 90%), and the predictions are not specific at the individual level.

3. Why do trainee GPs fail their exams[1]?

Speciality training is a difficult time for doctors, as they develop their clinical skills and increase their responsibilities. This may lead to an increase in stress, which in turn has negative effects on memory, attention and decision making. In addition, external factors such as workload, staff shortages, resource issues, performance measures and local management may impact negatively on performance, for reasons which are not the doctor's fault. All of these may impact on exam success, and in turn, repeated failure is a stressor in its own right. It is possible to identify from the international literature a number of construct irrelevant features which may lead to failure.

Six main contributing factors relevant to the present report emerged from the review conducted by Rothwell (2017)[xvi]. These were (a) having trained outside the country of practice combined with ethnicity, (b) gender, (c) age, (d) personality traits, (e) financial issues, and (f) social background.

Training outside the country of practice

Extensive evidence indicates that practicing in a different country from that in which medical training was undertaken is a significant challenge. For brevity, such doctors will be referred to as International Medical Graduates or IMGs.

[1] This section draws on work by Charlotte Rothwell, one of my PhD students, now published in a PhD Thesis from Durham University[xvi].

MacLellan[xvii] reported retrospectively on the success of IMGs who were pursuing or had completed a Quebec residency training programme and examinations. The success rates of IMGs (56%) were below that of Canadian and American graduates (93.5%).

Zulla et al (2008)[xviii] indicated that IMGs were less likely to join study groups during exam preparation. Crucially, sitting assessments too soon after joining the NHS was a significant factor in failing exams.

Esmail and Roberts (2013)[xix] analysed 5095 candidates who sat the AKT and CSA components of the MRCGP between 2010 and 2012. Black and Minority Ethnic (BME) IMGs were fifteen times more likely to fail the CSA at their first attempt than their white UK colleagues. These authors hypothesised that "Subjective bias due to racial discrimination in the clinical skills assessment may be a cause of failure for UK trained candidates and international medical graduates". However, the evidence that this results from systematic bias does not seem to be strong (Wakeford et al 2015[xx]; McManus et al, 2013[xxi]).

Atri et al (2001)[xxii] noted differences in interpretation of IMG difficulties. Supervisors thought that communication skills, relationships in the work place and working with other professionals were challenges to IMGs, while the IMGs themselves reported learning about the system and its values the most challenging aspects.

Broquet and Punwani (2012)[xxiii] found that that the understanding of feedback is culturally constructed. IMGs may view it as critical rather than supportive, and their trainers may in turn view their responses as indicating unwillingness to develop or respond.

Mahajan and Stark (2007)[xxiv] found that IMGs had a "'fear of losing face', which made them more inclined to hide or not accept their mistakes. As a result it was harder to learn from them". They also noted that "Lack of information about the National Health Service (NHS)/Royal Colleges, inappropriate communication skills, difficulties in team working, difficulties in preparing for Royal College examinations, visa and job hunting, and social and cultural isolation were identified as major barriers. Problems arose not only from difficulties with language but also from use of local and colloquial words, different accents and difficulty in communicating sensitive issues. Lack of understanding of role in teams and difficulties in working in multi-professional settings all contributed to the problems".

Mehdizadeh et al (2017)[xxv] found that doctors require to undertake performance assessment by the GMC varied significantly by place of medical qualification. Doctors from Bangladesh were 13 times more likely to have undergone performance assessment, followed by Nigerian and Egyptian trained doctors (eight times more likely). Doctors trained in Germany were over-represented amongst EEA graduates, who in turn were over four times more likely to undergo performance review. However, particularly for EEA graduates, 'place of training' can be confounded with 'ethnicity', as increasing numbers of UK citizens undertake medical training in mainland Europe.

The most recent Technical Report on the use of Situational Judgement Tests in the UK Foundation Programme Office (in press) illustrates this point well. Candidates self-identifying as 'White' do better than candidates identifying as BME amongst UK graduates.

But BME candidates trained in the UK do better than those white candidates who trained outside the UK, suggesting that ethnicity is subordinate to the cultural impact of the country of medical training in this instance.

Gender

In general, females outperform males in a wide variety of educational settings. As doctors, they have been found to be less likely to be referred for disciplinary action[xxvi], and have better patient outcomes (as well as costing less), as indicated earlier.

There may however, also be gender differences in factors affecting underperformance. In a longitudinal study Campbell et al (2010)[xxvii] surveyed internal medicine trainees over a three-year period (starting from their intern or first foundation programme year), using Maslach's Burnout Inventory (MBI). They found that persistent burnout was more likely to occur in males and had a positive association with depression in their internship year.

Coping strategies were found to be different in male and female residents (Lue et al, 2010)[xxviii]. Male residents were more likely to use a disengagement coping strategy than females. Emotion-focused engagement strategies were preferred by females. If trainees had more social support and utilised it, then they reported that they felt less tension-anxiety, depression-dejection and confusion-bewilderment. Females were found to be more likely to use social support as a coping strategy than male residents.

Females may encounter a number of stressors which are different from those of males (Rothwell, 2017 ibid). They may be more likely to work part time, and therefore feel that they are unfairly judged by comparison with those working full time. They may have primary child care responsibilities, and feel obliged to juggle work and life priorities in ways different to men.

Age

Younger trainees still making their way in their career may be more stressed, due to the regular social evaluative threats posed by assessment regimes, and by work life balance (Rothwell, 2017, ibid).

For IMGs, Zulla et al (2008)[xxix] found that they are often older and at a different life stage to those of their peers, having previously trained and worked in their own country then starting again in the training system in the host country.

Against this, Laidlaw et al (2006)[xxx] found that younger residents were better communicators than older ones. There is also some evidence that older doctors underperform compared to younger doctors with regard to patient outcomes (Norcini et al, 2014[vii]).

Personality Characteristics

Several studies reported specific personality traits which were linked with residents at risk of experiencing stress or burnout. This had a negative effect on patient care and may be an associated cause of depression in trainees.

The inability to recognise ones emotions can be linked to burnout and can potentially cause difficult interactions (Sargent et al, 2006)[xxxi]. Trainees who have low self-esteem, a 'victim mentality', low resilience, high neuroticism and low conscientiousness are more likely to suffer from stress and burnout (Hyman, 2011)[xxxii].

Financial Issues

Financial issues such as high levels of educational debt as significant stressors for doctors, contributing to stress and burnout (Rothwell, 2017 ibid). West et al (2011)[xxxiii] found that IMGs with high levels of debt were significantly more likely to suffer from stress and burnout. This debt may be a surrogate for other factors, such as having moved countries or coming from a different socio-economic background. Increasing student debt may also affect UK graduates.

Social Background

Social background is particularly interesting, since ethnicity may be confounded with social background. Yates (2010)[xxxiv] found that students from lower social classes were more likely to experience difficulty during the medical programme, including failing exams and late graduation. This effect persisted into post qualification experience, with such individuals being less likely to achieve consultant status.

Factors relating to Assessment

The above factors relate to the characteristics of those being assessed. However, there are further construct irrelevant factors relating to the assessment process itself. These are the nature of the assessment process itself, and the phenomenon of 'failure to fail'

Reliability and validity of assessment
As indicated in the 'Key concepts' section, no assessment process can separate True Positives from True Negatives. The proportions of each passing the assessment depend on the technical reliability of the assessment, and in a wider sense, its validity. Reliability is easiest to assess. Feinberg et al (2015) calculate the reliability of the score achieved in re-sitters in professional credentialing exams, and found that regression to the mean would account for increase of 0.14 Standard Deviations in re-sit score, due to measurement error in the assessment.
A much wider question, however, relates to the validity of the assessment process. Do the components of the AKT/CSA exams have predictive validity for future performance, and are they both *specific* and *sensitive*, in terms of a screening test? I am not aware of current data which quantifies these, and as a consequence, there is always a question about high stakes decisions made as a consequence of the current process.

'Failure to Fail'

It may seem counter-intuitive to consider the phenomenon of 'failure to fail' in the context of candidates who by definition have failed assessments! However, the question relates to whether or not they were correctly judged as ready to sit the AKT/CSA exams in the first instance.

'Failure to fail' (Cleland et al, 2008)[xxxv] refers to the situation when assessors who believe that a candidate ought to fail an assessment, nonetheless rate them as a pass. Reasons for this include concerns about the impact on the candidate, uncertainties about the value of the assessment, concerns about their own performance as a trainer or tutor, and perhaps most significant of all, the impact on themselves as assessors if they award a fail grade. Will the candidate complain about the outcome, perhaps raising issues of discrimination, of lack of 'fairness'? It is easier to assume that someone else along the line will fail them. In the setting of the AKT/CSA exams, this may lead to trainees not being guided that they are not ready to undertake the exams, when perhaps they require more training and preparation time. It would be valuable to have a process in place to ensure that candidates are signed off as ready to sit the exams. Of course, candidates may also have pressing reasons themselves for wishing to sit the exams as soon as possible, and may press for early dates for a variety of reasons.

Summary: reasons why trainees may fail assessments such as the AKT and CSA are complex and may include construct irrelevant factors. From the candidates' perspectives, these may include the interaction between country of training and ethnicity, gender, age, personality characteristics, financial circumstances, and socio-economic background. The assessment itself may be unreliable, and assessors may 'fail to fail' candidates in preliminary stages. Guidance might be offered to supervisors on determining when trainees are truly ready to sit, and a 'sign-off' of readiness to undertake the AKT/CSA exams introduced.

4. What is the relationship between re-sit performance, number of re-sits and the properties desired of those succeeding?

Postgraduate medical professional exams (for accreditation, Royal College Membership, and licencing purposes) are common. But deciding when candidates have achieved the desired standard is not always straightforward. This is important, as those who reach the cut score criteria generally gain significant benefit, in employability or legal entitlement to practice medicine, as well as demonstrating that they have attained a measurable standard of proficiency and undertaken personal development. Such assessments are therefore high stakes for doctors. But they are also high stakes for other parties, such as employers, colleagues, and, crucially, patients.

All testing methods are imprecise, and there is always an error component in all scores obtained. Therefore there is a risk of false negatives (those who fail, but deserve to pass), and false positives (those who pass but who deserve to fail). There is a significant body of evidence that indicates that candidates do better on re-sit than they do on their first attempt (Matton et al, 2009)[xxxvi]. Those who fail on a first or subsequent sitting may be able

to re-sit again but there are significant cost implications. The individual generally has to pay for the assessment, and will lose income through lack of career progression. However, there are also opportunity costs to the community through (a) being deprived of a doctor during the re-assessment period, or by (b) supporting further time in training. For example, this is estimated at £40, 000 for an extra six months in UK general practice[xxxvii] . False positives, on the other hand, pose potential hazards to patients, resulting in expensive resolution processes, and may have knock on consequences for employers who are exposed to an increase in legal liability.

Re-sit methodology is therefore important in all medical jurisdictions across the world. However there are only a limited number of practical and theoretical studies of the consequences of allowing re-sits on test performance, with commentators remarking on the dearth of evidence (Cohen-Schotanus, 1999)[xxxviii]. For example, Ricketts (2010)[xxxix] commented that "there is no 'theory of re-sits' " and "there is much common practice but no evidence base for the interpretation of re-sit results."

I have reviewed the evidence on the number of re-sits appropriate for a high stakes professional medical assessment and included studies on subjects other than medicine , as studies focussing purely on doctors are limited in number.

Scores increase with re-sitting: non-medical findings

In general, candidate scores increase with re-sits (reviewed in Matton et al, 2009)[xl].

Multiple form testing

Studies may use *identical* forms (the same test given again after an interval), *parallel* forms (a similar test but with different items), and *different* forms (a different kind of test, aimed at the same construct). When Air Force reservists were given either an identical form, parallel form, or different form test, 10 minutes or 7 hours apart, (Krumboltz and Christal, 1960)[xli] in a spatial aptitude test there were gains of the order of 1 SD observed for the first two forms but not for the different form, which they interpreted as indicating that the gains on identical and parallel tests are construct irrelevant. However in the US Department of Labor General Aptitude Test (which tested cognitive ability and physical dexterity) re-testing after two weeks, with one sub-group sitting an identical test, and the other a parallel test (United States Department of Labor, 1970)[xlii]. demonstrated effect sizes for the identical test of 0.32 – 0.74 and for the parallel test form, of 0.15 – 0.55.

Re-sitters

An important meta-analysis (Kulik et al, 1984)[xliii] looking at re-sitters (repeat testing for those who had previously failed) indicated that test scores increased by 0.42 SD at the second administration of an identical test, and by 0.23 SD at the second administration of a parallel test. But interestingly, these authors also found a significant positive relationship between ability and the score increase observed on re-take. High ability re-sitters (i.e. those just below the cut score) had a score increase of 0.80 SD, middle ability re-sitters of 0.40 SD and low ability re-sitters of 0.17 SD. They also observed a 'dose response' effect of multiple

re-sittings, with SD gains of 0.42 from 1st to 2nd re-sit, 0.70 from 1st to 3rd re-sit, and 0.96 from 1st to 4th re-sit. These data, however, relate to identical test forms, and may not correspond to other related test forms.

Impact of feedback & targeted learning
One study (Friedman, 1987)[xliv] explored the impact of feedback and targeted learning on resitting an introductory statistics course. He found a significant increase in scores with over 90% of individuals improving their personal score, indicating that the effect was consistent and uni-directional. Similarly other studies show that students who re-sit to improve scores rather than to pass after a previous failure, do better following further study (Cates 2001; Juhler 1998)[xlv],[xlvi].

Scores do not increase continuously
From the literature there is good evidence that candidate performance does not increase consistently or continuously over each attempt. For example, a study of applicants to a law enforcement programme (Hausknecht et al 2002)[xx] noted that performance increased significantly on the second and third attempts, but showed no further gain on the fourth attempt. Geving et al (2005)[xlvii] reviewed literature on re-sit performance, summarising as a key finding that repeated exposure to items promotes score increases beyond those of latent trait change. They concluded that test scores increased but "the number of retakes and score gains were inversely related, indicating that, after the second testing opportunity, score gains were not as great". Previous exposure to items did not seem to have a significant effect, but length of time between test attempts did in that it was positively related to scores, suggesting learning had taken place.

In a meta-analysis, Hausknecht et al (2007)[xlviii] found an overall increase of about 0.25 SD on re-sit, based on 107 studies of cognitive oriented tests. Schleicher et al (2010)[xlix] reported a re-test effect of 0.15 SD on a job-knowledge test given to federal agency job applicants.

Many of the above results relate to cognitive ability tests. As we will see, results may be different for declarative knowledge tests.

Studies Focused on Medical Settings

Findings from medical settings are similar to those obtained in non-medical settings. Since this is a field in development, studies are summarised in chronological order.

In 1992, McManus indicated that successful re-sit candidates in the MRCGP examination do not solely pass the first re-sit on the basis of increased performance; having different questions they can answer seems to be more significant. However, they do improve in performance on their second and third re-sit[l] indicating that more targeted learning has taken place and suggesting that additional time to enable further learning is important.

Bandaranayake and Buzzard (1994)[li] noted that, with regard to the Royal Australian College of Surgeons Part 1 exam, the probability of re-sit candidates passing remained fairly

constant up to the fourth attempt and fell thereafter. The lower the candidates' original mark had been, the less likely they were to pass on subsequent attempts.

Boulet et al (2003)[lii] investigated the performance of first time and repeat candidates in a high stakes, high fidelity, standardised patient assessment, using both identical form and parallel form assessments. These were candidates who were undertaking the Educational Commission for Foreign Medical Graduates (ECFMG) Clinical Skills Assessment (CSA®). There were significant (p<.01) increases in candidate scores between the first and second attempts (undertaken within six months of the initial CSA) for all of the CSA components. There was no difference between identical and parallel test forms, except that non-US international medical graduates did slightly worse when they were exposed to repeat information.

A study of candidates for medical school in Belgium (Lievens et al, 2005)[xii] showed that candidates performed significantly better on re-sitting knowledge, situational judgement and cognitive ability tests, with the improvement being most marked in the last of these[liii]. Subsequent performance was also explored. For those who passed first time, higher performance was associated with higher knowledge scores. For those who passed on re-sit, subsequent performance was associated with their re-sit score rather than their initial fail score. Hence, for these students, the second attempt had higher validity than the first attempt. Conversely, cognitive ability tests showed the opposite effect, suggesting that in this case, score improvement was construct irrelevant.

Reiter et al (2006)[liv] showed that advance access to Multiple Mini Interview questions through security violations did not lead to significant increases in score.

Hays et al (2008)[lv] explored re-sit performance by degree of severity of failure, banded by factors of Standard Error of Measurement (SEM) and found, as might be expected, that candidates who missed the pass mark by 1 SEM had a reasonable probability (83%) of passing on the second attempt: those who failed by 3 SEM had 100% probability of failing the re-sit or withdrawing.

Griffin et al (2008)[lvi] explored the effect of coaching and re-testing on multiple Mini Interview (MMI scores) and UMAT scores. Re-testing on MMI had no significant benefit for new stations, but had a small benefit for repeated or very similar stations, suggesting that any improvement was due to construct irrelevant familiarity. Coaching had no significant benefits, and indeed, scores on one station were significantly reduced in the coached group. Coaching also had no benefit for UMAT, but re-testing was not examined for this test. MMIs are similar to OSCEs, and therefore information from this source may be relevant to OSCE performance.

In a study by Raymond et al (2009) [lvii] looking at re-test effects on credentialing exams for radiographers, 541 examinees who had previously failed a national certification exam on their first attempt were randomly assigned to receive either the same paper again or a parallel paper during their second attempt. The study found that although the group who had received the same paper had a shorter response time, the mean scores for the paper were not higher.

Raymond and Luciw-Dubas (2010)[lviii], in studies on candidates for medical speciality boards in the US, found that re-sitters improved scores more on oral exams than on written tests, and question whether such improvement is due to construct irrelevant factors. Note that oral tests may be closer to ability tests than knowledge tests, however.

A study by Swygert et al (2010)[lix] investigated gains for repeat examinees, where they had experienced repeat information, for the USMLE Step 2 Clinical Skills exam. This is a scenario based assessment where candidates interact as doctors with standardised patients. A large data set (n=3045) of candidates who had failed their initial exam were retested. They found that there was a significant score increase in their second attempt in all four areas of the Step 2 CSA. However they observed no significant difference in candidates who had previous exposure to the exam information.

Raymond et al (2011)[lx] reviewed performance of re-takers on USMLE Step 2. This showed that first time failures had a markedly different factor structure than first time passers, but on their second attempt became more like first time passers. Comparison with subsequent clinically related performance showed that the re-sit score had more validity than the initial score, in findings similar to those of Lievens et al (2005)[lxi] for knowledge tests.

Raymond et al (2012)[lxii] analysed scores for single-take examinees and repeat examinees who completed a 6-hour clinical skills assessment required for physician licensure. Each examinee was rated in four skill domains: data gathering, communication-interpersonal skills, spoken English proficiency, and documentation proficiency. They concluded that it is valid to draw inferences from re-sit scores.

Probably the most in depth analysis of re-sit performance published so far on the impact of multiple re-sits on test scores is that of McManus and Ludka (2012)[lxiii]. These authors analyse attempts from 2002/3 to 2010 at the Royal College of Physicians (UK) Membership tests, which are in three parts. Parts 1 and 2 are written (Best of Five MCQs): Part 3 (Practical Assessment of Clinical Examination Skills) is in the form of a practical multi station OSCE-style test, but using real patients. Since the number of attempts has been unlimited, data is available on the consequences on candidate scores of large numbers of re-sits (with two candidates sitting Part 1 26 times, and one candidate requiring 35 attempts to pass all three Parts). Limitations include that the data is left and right truncated in that some candidates had sat the various parts before the study start date, and some candidates who had failed at the study end date would undertake further re-sits, but these are not likely to be significant factors. Using a variety of models, the authors suggest that scores increased up to the tenth attempt at Part 1, the fourth attempt at Part 2, and the sixth attempt at PACES. As a general conclusion, the authors suggest that there is no clear rational basis for limiting the number of re-sits. They suggest the possibility of increasing the cut score for candidates at each re-sit (along the lines suggested by Millman (1989)[lxiv].

Chavez et al (2013)[lxv] explored whether improved performance on re-sits on a standadised patient examination was due to familiarity. They found that performance improved after the first few stations, indicating that candidates took a little while to 'calibrate' for the stations. However, this did not carry over into the second sitting, where the same

phenomenon was observed. As they conclude: "The within-session score gains over the first three to six SP encounters of both attempts indicate that there is a temporary "warm-up" effect on performance that "resets" between attempts. Across-session gains are not due to this warm-up effect and likely reflect true improvement in performance".
Pell et al (2012) [lxvi] unusually showed evidence that the performance of re-sit OSCE candidates declined across repeated attempts, despite conscious efforts at remediation.

A most informative study is that of Feinberg et al (2015)[lxvii]. They confirm the apparently paradoxical findings that re-sitters taking an identical form test of knowledge again do not improve their performance due to familiarity with the material. On the contrary, they do just as badly on individual items as they did before. Other candidates sitting a parallel form of the test improve their performance, because different questions allow them to demonstrate other areas of expertise.

Tiffin et al (2017)[lxviii] found that low scores on the GMC Professional and Linguistic Board examinations (both written and practical) were significant predictors of the likelihood of later censure by the GMC, as were the number of re-sits undertaken (without a requirement for further training).

Summary: Scores generally improve on re-sit, by amounts typically of 0.3 or 0.4 Standard Deviations, but the improvement decreases with each attempt, and may plateau after two or three attempts. Scores can improve on re-sitting due to three factors: familiarity with test material, statistical variation, and improvement on the construct under test. Short tests of ability show most effect of familiarity; improvement on longer credentialing or achievement tests suggests that there is improvement on the test construct. Candidates on longer achievement tests do not receive an advantage from sitting the same test items again. Candidates improve on the test construct after further study.

5. Risks associated with passing poor performers

As indicated above, calculating the respective costs of False Positives and False Negatives is a cost benefit analysis. I know of no data in which the cost ratio has been calculated. However, the error rate of a non-existent doctor, who therefore makes no decisions, is likely to be higher than the error rate of an existent doctor, we may hope. The observation that death rates may stay the same or even fall during doctors' strikes is probably due to the delay or cancellation of high risk procedures[lxix].

A 2012 estimate suggested that there are some 12, 000 avoidable deaths per year in the NHS (Hogan et al, 2012)[lxx]. NHS England reported 306 'Never Events' (serious issues such as patient misidentification, leaving instruments behind, or operating on the wrong organ) in 2014-15 (NHS England, 2016)[lxxi]. There were 8884 complaints about doctors to the GMC in 2014 (GMC 2015)[lxxii], a slight fall against the previous year, but against a generally rising trend.

Donaldson et al (2014)[ii] found that the annual referral rate to the National Clinical Assessment Service was approximately five per 1000 doctors (95% CI 4.6 to 5.4), but cited earlier work indicating that over a five year period, 6% of doctors caused concern. Doctors whose first medical qualification was gained outside the UK were more than twice as likely to be referred as UK-qualified doctors; male doctors were more than twice as likely to be referred as female doctors; and doctors in the late stages of their career were nearly six times as likely to be referred as early career doctors. The National Clinical Assessment Service reported 175 doctor (including GP) suspensions in 2013-4 (NCAS Report 2016)[lxxiii]. However, 44% of these recent NCAS cases related to non-UK graduates and 56% related to UK graduates, meaning that non-UK graduates are significantly over-represented in these recent figures, as in the Donaldson et al study [ii].

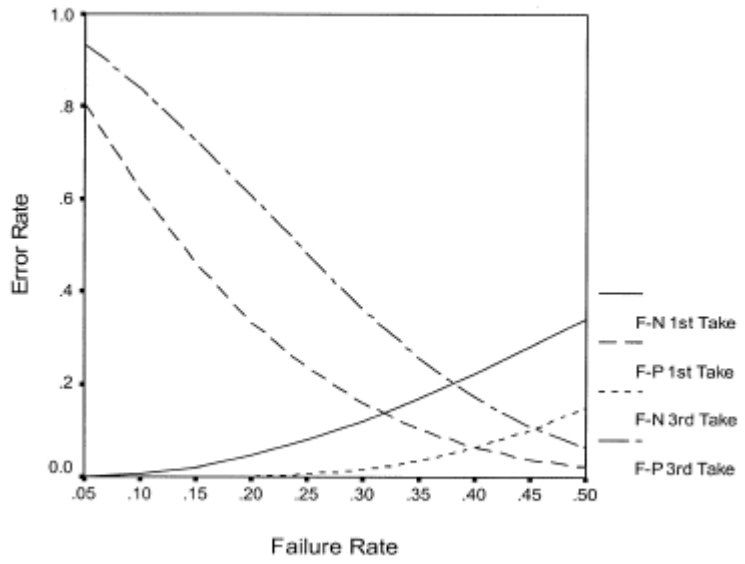
In the absence of economic data, a number of papers have explored *simulations* of exam performance under various testing conditions. The advantage of such studies is that they set aside all factors other than exam reliability – in other words, neglecting factors such as candidate health and wellbeing, or the details of the testing environment.

For example, Millman (1989)[lxxiv] presented evidence of the impact of repeat testing on candidates of varying proficiency. This indicated that candidates whose True Score was 5% above the cut score representing proficiency had a 90% chance of passing first time (hence with a 10% false negative rate). Borderline candidates whose true score equals the cut score have a better than 90% chance of passing after three re-sits. But candidates whose True Score is 5% below the proficiency cut score have a 15% chance of passing on the first attempt, and an 80% chance of passing after 10 attempts.

Clauser and Nungster (2001)[lxxv] demonstrated that a simulated test with a reliability of 0.92, applied to candidates, 90% of whom are proficient, would have a false positive rate of 20%, doubling after two re-sits, and at that point corresponding to the false positive rate of a test with a reliability of 0.69. As these authors indicate, false positive errors "may put the public at risk by allowing unqualified candidates to become licensed or certified". The false positive rate increases as reliability decreases, but decreases as the cut score increases. They suggest that "One important strategy is to limit the number of retakes". Another is to increase the initial cut score, especially where the cost of a false positive is higher than that of a false negative. Finally, the paper explores the consequence of raising the cut score for re-sits, as suggested by Millman (1989)[lxxvi]. The effect of increasing the cut score by 0.25 SD per administration is considered, along with possible resistance to this approach.

Further theoretical models (Clauser et al, 2006)[lxxvii] indicated clearly that if either the cut score is reduced **or** the number of re-sits increased in an effort to reduce false negatives, then the false positives increase disproportionately (Figure 1).

Figure 1 (from Clauser et al, 2006). The conditional false positive (FP) and false negative (FN) rates for a single administration of a test and for three administrations (i.e. 2 re-sits) of the same test



This is partly because the situation is asymmetric. If a candidate fails, they are allowed to sit again: if they pass, they are not required to sit again, even though there will be false positives amongst those passing. The costs of this are also of interest, and depend on (a) the cost to the individual of a false negative, (b) the benefit to the individual of a false negative (they have to sit again, and may learn more), (c) the cost to society of a false positive (risk of poor medical treatment) and (d) the cost to society of a false negative (if doctors are in significant under-supply then a less strong doctor may be better than none) (Figure 2).

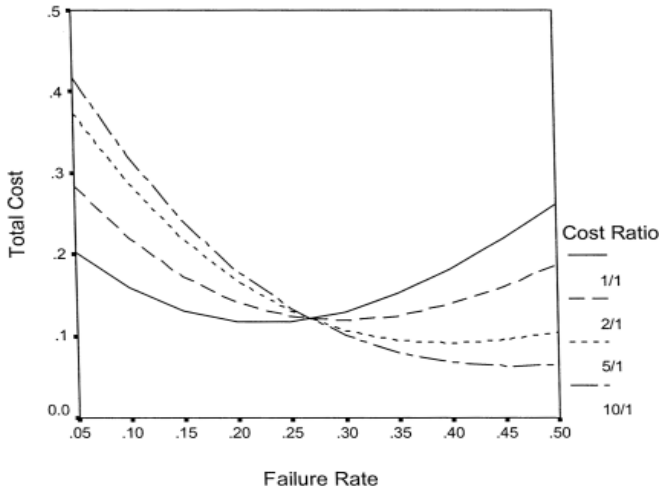


Figure 2 (from Clauser et al, 2006). Costs from false positives increase as failure rates decrease, and the ratio of false positive costs to false negative costs rises.

Tighe et al (2010)[lxxviii] describe a Monte Carlo Simulation in which theoretical candidates sit a high reliability test, and then sit it again. While some 16% pass on the first attempt, only 11% pass on both occasions. However, by setting the cut score at 60%, where the mean score is 50%, this simulation increases the number of 'false negatives'. Not many professional exams consistently have a fail rate of 85% as in this version of the model, and this would probably be viewed as a cause for concern in reality.

There are potentially serious cost consequences of both false positives and false negatives in high stakes medical assessments, although the strategy of testing bodies may be influenced by the fact that the cost of false positives falls largely on others, whilst challenges arising from false negatives may fall on the testing body itself. The total societal cost depends on the ratio of false positives to negatives, the respective costs of each, and the relative likelihood of each under different assessment regimes. The assessment regime in this context includes the standard setting process for the assessment, the validity and reliability of the assessment, and the opportunities for subsequent re-sits.

If it is accepted that false positives arising from unlimited re-sit opportunities are a significant hazard, the literature suggests three ways of dealing with this. The first is to limit the number of re-sits, the second is to increase the cut score on re-sit attempts as opposed to first attempts (Millman, 1989[lxvi]; McManus and Ludka, 2012) , and the third is to employ an averaging process for re-sit scores (Millman, 1989[lxvi]).

While there are psychometric arguments for each of these, there may be differences in acceptability to the candidate groups. Increasing the cut score for re-sits is likely to face challenges on the basis of strict egalitarianism, since some re-sit candidates would fail with a score which will pass first-sit candidates sitting exactly the same test at the same time. Similarly, averaging scores across occasions will mean that some candidates fail because of their performance on a previous test rather than the current test. I believe that limiting the number of re-sits is likely to be perceived as the most just method of achieving the aim of reducing the number of false positives to an acceptable level, and offers the opportunity for more targeted learning between tests.

The justice of this approach will be even more clearly evident if a 'refractory period' is invoked rather than a blanket ban on ever sitting the assessment again. This refractory period might include a requirement for structured training of some kind, or the passage of sufficient time in practice for additional relevant clinical experience to be gained.

If this approach is to be followed, a key question is obviously 'how many re-sits is enough' in any particular setting. Obviously, this could be established on a case by case basis, by study of the performance of any given assessment in terms of its reliability, validity and cut score. However, while the last two of these can fairly readily be ascertained by psychometric analysis of past data, validity is harder to confirm. However, the evidence from the literature seems to suggest a possible generic answer. The evidence indicates that four attempts (one initial attempt and three re-sits) is a reasonable compromise in many circumstances (also recommended by McManus 1992)[xxxii]. The differences between four attempts and, say, six attempts as mentioned in McManus and Ludka (2012)[lv] are small, and probably not meaningful.

There is evidence to suggest that a refractory period before further attempts would increase performance. There are well known effects of the passage of time on psychomotor skills (Caretta et al, 2000)[lxxix], but I do not know of equivalent studies on professional examinations. For instance, candidates may improve with further experience, or deteriorate through lack of recent experience, or decline in performance with age. However, I believe that a refractory period of at least one year, but more probably two years allows sufficient

time for further development and improvement of performance, especially if associated with targeted feedback and training.

Summary: Both false positives and false negatives in professional medical exams have costs, both to the individual and society. False positives pose risks to patient safety, to the extent that safety errors are caused by individual errors rather than system factors. False negatives deprive society of needed doctors in times of shortage (and in the context of this report, deprive the NHS and public of GPs).

A limit is recommended of four attempts. Without re-training, passes after more attempts are likely to be false positives. However, improvement on scores, following further training, is likely to represent a true improvement.

6. Medical Error and Safety

There is an assumption running through these studies that patient risk is associated with the performance of the individual. Safety studies do not generally accord with this view. Errors are generally errors of the system[lxxx], although the individual doctor is the 'politically most blameable unit'. Factors such as work place management, staffing and lack of resources are also important. A culture of personal blame, in which responsibility is assigned to the individual, inhibits reporting of 'near misses', and removes the possibility of responding to latent errors in the system. Even for those circumstance where the individual doctor is at fault, then it needs to be clearly established that more stringent testing of knowledge and skills during training would have helped, given that most 'Fitness to Practice' issues relate to professional behaviours and attitudes, rather than straightforward knowledge and skills. It is a plausible assumption that reducing the number of doctors who qualify for a particular profession in itself is a challenge to workplace safety, since qualified doctors may then not be available, or available doctors may be over-stressed, and thus prone to mistakes. A vicious circle may be created, in which under-staffed working environments encourage doctors to leave.

In general, a distinction can be drawn between selecting environments (where there are many more applicants than places) and recruiting environments (where there are more places than applicants). It has been customary to regard medical employment as a selecting environment, where there is an excess of applicants. None the less, this is only true in certain specialities. And in the light of the increased demands for GPs (and psychiatrists), and the uncertainties about the availability of European Economic Area (EEA) trained doctors subsequent to Brexit, these customary expectations must be re-evaluated.

Summary: Medical error and patient safety are not simple matters of individual errors by doctors, which can be addressed just by raising barriers in medical professional exams. In particular, allowing fewer doctors to qualify for a profession or speciality raises the risk of under-staffing, which is in itself a major source of medical error.

7. Re-training and development

Obviously, mandating a gap before allowing a second set of re-sits would be purposeless without appropriate personal development in that time, and it is of interest to consider what that personal development programme might look like. Targeted reflection on the reasons for failure on the initial failure would be valuable, followed by actions to remedy weaknesses identified by the process. This depends on the degree of feedback that is provided to unsuccessful candidates by the RCGP. A recent study by one of my students[lxxxi] suggests that mentoring, simulation training, cultural awareness training, creating a 'buddying' system and awareness training for supervisors (including understanding the challenges of being an IMG, and reconciling work and family life) can all help improve performance and reduce medical error. Communication skills, and giving and receiving feedback in a positive spirit are all crucial. Methods from Crew Resource Management approaches (particularly SBAR - an acronym for Situation, Background, Assessment, Recommendation) and use of checklists may be more important contributors to safety than cut scores on professional assessments.

Summary: Receipt of structured feedback and targeted training for candidates who have failed during a mandated 'refractory' period is essential to improvement on the property under test.

8. Conclusions and Recommendations

Current evidence shows that the number of permitted re-sits should be limited, since those who have not passed after three or four attempts are not likely to pass on an immediately subsequent occasion. Perversely, permitting further attempts after this number is likely to pass only True Negatives, who pass through chance, test familiarity and the variability of the exam. However, there is clear evidence that performance can improve on further training and feedback. The optimal strategy would therefore be limit the number of re-sits to no more than four, then to require a mandatory period of targeted training of at least one year, and more plausibly two, after which a further limited number of re-sits should be permitted. By the same logic, the number of further attempts should also be limited, to a maximum of two, since candidates should be able to demonstrate that significant improvement has taken place since the last exam in a maximum of two attempts. The reason for choosing two further attempts, rather than one, is that for candidates who pass on a second attempt have a factor structure similar to those who pass on a first attempt.

Concerns may be expressed that this may allow doctors to pass who are not safe to do so. However, performance improvement subsequent to further training in an extended credentialing exam is evidence of an improvement in the construct under test. In any case, the numbers involved are likely to be small compared to the False Positives who passed the initial 'first sit' assessment.

In summary, the current RCGP process allowing multiple re-sits with no time restrictions, followed by a ban on further attempts is not optimal in terms of identifying those who

ought to pass or fail. It would be better to allow a maximum of four attempts, followed by a mandatory re-training period before allowing further attempts.

Attention should be given to the preparedness of candidates to undertake the AKT/CSA exams in the first instance, particularly with regard to the level of cultural familiarity of International Medical Graduates with practice in the UK.

References

[i] Firth-Cozens J. (2008) Doctors with difficulties: why so few women? *Postgrad Med J*. **84**:318-20

[ii] Donaldson L, Panesar SS, McAvoy P Scarrot DM (2014). Identification of poor performance in a national medical workforce over 11 years: an observational study. *BMJ Quality and Safety*. **23**:147-152

[iii] Tsugawa, Y., Jena, A.B., Figueroa, J.F., Orav, E.J., Blumenthal, D.M. and Jha, A.K.(2017). Comparison of hospital mortality and readmission rates for Medicare patients treated by male vs female physicians. *JAMA internal medicine*, **177**:206-213.

[iv] Sasso, A.T.L., Richards, M.R., Chou, C.F. and Gerber, S.E., 2011. The $16,819 pay gap for newly trained physicians: the unexplained trend of men earning more than women. *Health Affairs*, **30**:193-201.

[v] Hambleton, R.K. and Cook, L.L., 1977. Latent trait models and their use in the analysis of educational test data. *Journal of educational measurement*, **14**:75-96.

[vi] Hamdy H, Prasad K, Anderson MB, Scherpbier A, Williams R, Zwierstra R, Cuddihy H. (2006) BEME systematic Review: predictive values of measurements obtained in medical schools and future performance in medical practice. *Medical Teacher* **28**: 103-106

[vii] Tamblyn R, Abrahamowicz M, Dauphinee WD, Hanley JA, Norcin J, Girard N, Grand'Maison P, Brailovsky C (2002) Association between licensure examination scores and practice in primary care. *JAMA;* **288**: 3019-3026

[viii] Tamblyn R, Abrahamowicz M, Dauphinee D, et al. (2007) Physician scores on a national clinical skills examination as predictors of complaints to medical regulatory authorities. *JAMA* 298: 993–1001

[ix] Papadakis MA, Arnold GK, Blank LL, Holmboe ES, Lipner RS. (2008) Performance during Internal Medicine Residency Training and Subsequent Disciplinary Action by State Licensing Boards. *Annals of Internal Medicine* **148**:869-876

[x] Holmboe ES, Wang Y, Meehan T, et al. (2008) Association between maintenance of certification examination scores and quality of care for medicare beneficiaries. *Archives of Internal Medicine* **168**: 1396-403

[xi] Wenghofer E, Klass D, Abrahamowicz M, Dauphinee D, Jacques A, Smee S, Blackmore D, Winslade N, Reidel K, Bartman I, Tamblyn R. (2009) Doctor scores on national qualifying examinations predict quality of care in future practice. *Medical Education*; 43: 1166-1173

[xii] Norcini, J.J., Boulet, J.R., Opalek, A. and Dauphinee, W.D., (2014). The relationship between licensing examination performance and the outcomes of care by international medical school graduates. *Academic Medicine*, **89**:1157-1162.

[xiii] Tiffin, P. A., Illing, J., Kasim, A. S. & McLachlan, J. C**.** (2014). Annual Review of Competence Progression (ARCP) performance of doctors who passed Professional and Linguistic

Assessments Board (PLAB) tests compared with UK medical graduates: national data linkage study. *BMJ* **348**: g2622

[xiv] Tiffin, P.A., Illing, J., Kasim, A.S. and McLachlan, J.C., (2014). Annual Review of Competence Progression (ARCP) performance of doctors who passed Professional and Linguistic Assessments Board (PLAB) tests compared with UK medical graduates: national data linkage study. *BMJ*, **348**:g2622.

[xv] Archer, J., Lynn, N., Coombes, L., Roberts, M., Gale, T., Price, T. and de Bere, S.R., 2016. The impact of large scale licensing examinations in highly developed countries: a systematic review. *BMC Medical Education*, **16**:212.

[xvi] Rothwell C (2017) A study to identify the factors that either facilitate or hinder medical specialty trainees in their Annual Review of Competency Progression (ARCP), with a focus on adverse ARCP outcomes. Durham University PhD Thesis, approved for publication but not yet published.

[xvii] MacLellan AM. (2010) Examination outcomes for international medical graduates pursuing or completing family medicine residency training in Quebec. *Canadian Family Physician*. **56**: 912-919

[xviii] Zulla, R., Baerlocher, M.O. and Verma, S., 2008. International medical graduates (IMGs) needs assessment study: comparison between current IMG trainees and program directors. *BMC Medical Education*, **8***:*42.

[xix] Esmail A and Roberts C. Academic Performance of ethnic minority candidates and discrimination in the MRCGP examinations between 2010 and 2012: analysis of data. *BMJ* **347**: f5662.

[xx] Wakeford, R., Denney, M., Ludka-Stempien, K., Dacre, J. and McManus, I.C., 2015. Cross-comparison of MRCGP & MRCP (UK) in a database linkage study of 2,284 candidates taking both examinations: assessment of validity and differential performance by ethnicity. *BMC medical education*, **15**:1.

[xxi] McManus, I.C., Elder, A.T. and Dacre, J., 2013. Investigating possible ethnicity and sex bias in clinical examiners: an analysis of data from the MRCP (UK) PACES and nPACES examinations. *BMC medical education*, **13***:*103.

[xxii] Atri A, Matorin A, Ruiz P. (2001) Integration of International Medical Graduates in US psychiatry: the role of acculturation and social support. *Academic Psychiatry* **35**:21-25

[xxiii] Broquet K, Punwani M. (2012) Helping international medical graduates engage in effective feedback. *Academic Psychiatry* **36**:4

[xxiv] Mahajan J. Stark P (2007) Barriers to education of overseas doctors in paediatrics: a qualitative study in South Yorkshire. *Arch Dis Child*. **92**: 219-223.

[xxv] Mehdizadeh, L., Potts, H.W.W., Sturrock, A. and Dacre, J., (2017). Prevalence of GMC performance assessments in the United Kingdom: a retrospective cohort analysis by country of medical qualification. *BMC Medical Education*, **17**:67.

[xxvi] Firth-Cozens J.(2008) Doctors with difficulties: why so few women? *Postgrad Med J*. **84**:318-20

[xxvii] Campbell J, Prochazka AV, Yamashita T, Gopal R. (2010) Predictors of persistent burnout in internal medicine residents: a prospective cohort study. *Academic Medicine* **85**:1630-4

[xxviii] Lue, B.H., Chen, H.J., Wang, C.W., Cheng, Y. and Chen, M.C., (2010). Stress, personal characteristics and burnout among first postgraduate year residents: a nationwide study in Taiwan. *Medical teacher*, **32**:400-407.

[xxix] Zulla R, Baerlocher MO, Verma S. (2008) International Medical Graduates (IMGs) needs assessment study: comparison between IMG trainees and program directors. *BMC Medical Education*. **8**:42.

[xxx] Laidlaw TS, Kaufman DM, MacLeod H, van Zanten S, Simpson D, Wrixton W. (2006) Relationship of resident characteristics, attitudes, prior training and clinical knowledge to communication skills performance. *Medical Education* **40**: 18-25.

[xxxi] Sargent C. Sotile W, Sotile M, Rubash H, Barrack RL. (2004) Stress and coping among orthopaedic surgery residents and faculty. *Journal of Bone and Joint Surgery*. **86**:1579-1586

[xxxii] Hyman SA. (2011) Risk of burnout in perioperative clinicians: a survey study and literature review. *Anaesthesiology*. **114**: 194-204

[xxxiii] West CP, Shanafelt TD, Kolars JC. (2011) Quality of life, burnout, educational debt, and medical knowledge among internal medicine residents. *JAMA* **306**:952-960.

[xxxiv] Yates J, James D (2010) Risk factors at medical school for subsequent professional misconduct: multicentre retrospective case-control study. *BMJ* **340**;c2040

[xxxv] Cleland, J.A., Knight, L.V., Rees, C.E., Tracey, S. and Bond, C.M., (2008). Is it me or is it them? Factors that influence the passing of underperforming students. *Medical Education*, **42**:800-809.

[xxxvi] Matton N, Vautier S, Raufaste E. (2009) Situational effects may account for gain scores in cognitive ability testing: A longitudinal SEM approach. *Intelligence;* **37:** 412-21

[xxxvii] Crawford ME(2005) Reassuring evidence on competency based selection. *BMJ* **330**:711-4

[xxxviii] Cohen-Schotanus, J.A.N.K.E., (1999). Student assessment and examination rules. *Medical Teacher*, **21**:318-321.

[xxxix] Ricketts, C., 2010. A new look at resits: are they simply a second chance?. *Assessment & Evaluation in Higher Education*, *35*(4), pp.351-356.

[xl] Matton N, Vautier S, Raufaste E. (2009) Situational effects may account for gain scores in cognitive ability testing: A longitudinal SEM approach. *Intelligence;* **37:** 412-21

[xli] Krumboltz, J. D. & Christal, R. E. (1960). Short-term practice effects on tests of spatial aptitude. *Personnel and Guidance Journal,* **38**, 385-391.

[xlii] United States Department of Labor. (1970) *Manual for the USES General Aptitude Test Battery*. Washington, D.C.: U.S. Department of Labor.

[xliii] Kulik JA, Kulik CC, Bangert RL. (1984) Effects of practice on aptitude and achievement test scores. *American Educational Research Journal;* **21***:* 435–447

[xliv] Friedman, H. (1987). Repeat examinations in introductory statistics courses. *Teaching of Psychology,* **14**: 20-23.

[xlv] Cates, W. M. (2001). The efficacy of retesting in relation to improved test performance of college undergraduates. *Journal of Educational Research,* **75**:230-236.

[xlvi] Juhler, S. M., Rech, J. F., From, J. F., & Brogan, M. M. (1998). The effect of optional retesting on college students' achievement in an individualized algebra course. *Journal of Experimental Education,* **66**:125-138.

[xlvii] Geving AM, Webb S et al. (2005) Opportunities for repeat testing: Practice doesn't always make perfect. *Applied H.R.M. Research;* **2**:47-56

[xlviii] Hausknecht JP, Trevor CO, Farr JL. (2002) Retaking ability tests in a selection setting: Implications for practice effects, training performance, and turnover. *Journal of Applied Psychology;* **87:** 243-254

[xlix] Schleicher DJ, Van Iddekinge CH, Morgeson FP, et al. (2010) If at first you don't succeed, try, try again: understanding race, age, and gender differences in retesting score improvement. *Journal of Applied Psychology;* **95:** 603-617

[l] McManus IC, Lockwood DNJ. (1992) Does performance improve when candidates resit a post-graduate examination? *Medical Education;* **26**: 157-162

[li] Bandaranayake RC, Buzzard AJ. (1993) The probability of passing at resits in the part 1 fellowship examination. *Australian and New Zealand Journal of Surgery;* **63**: 723-726

[lii] Boulet JR, McKinley DW, Whelan GP, Hambleton RK. (2003) The effect of task exposure on repeat candidate scores in a high-stakes standardized patient assessment. *Teaching and Learning in Medicine;* **15**:227-232

[liii] Lievens F, Buyse T, Sackett PR. (2005) Retest effects in operational selection settings: Development and test of a framework. *Personnel Psychology;* **58:** 981-1007

[liv] Reiter HI, Salvatori P, Rosenfeld J et al (2006) The effect of defined violations of security on admissions outcomes using multiple mini interviews. Medical Education **40**: 36-42.

[lv] Hays RB, Sen Gupta TK, Veitch J. (2008) The practical value of the Standard Error of Measurement in borderline pass/fail decisions. *Medical Education*; **42**: 810–815

[lvi] Griffin B, Harding WH, Wilson G, Yeomans ND (2008) Does practice make perfect? The effect of coaching and re-testing on selection tests used for admission to an Australian Medical School. *Medical Journal Australia* **189**: 270-273.

[lvii] Raymond MR, Neustel S, Anderson D. (2009) Same-Form retest effects on credentialling examinations. *Educational Measurement: Issues and Practice*, **28**: 19-27.

[lviii] Raymond, M.R. and Luciw-Dubas, U.A., 2010. The second time around: accounting for retest effects on oral examinations. *Evaluation & the health professions*, **33**:386-403

[lix] Swygert KA, Balog MS, Jobe A. (2010) The Impact of Repeat Information on Examinee Performance for a Large-Scale Standardized-Patient Examination. *Academic Medicine;* **89**:1506-1510

[lx] Raymond MR, Kahraman N, Swygert KA, Balog KP. (2011) Evaluating Construct Equivalence and Criterion-Related Validity for Repeat Examinees on a Standardized Patient Examination. *Academic Medicine;* **86**: 1253-1259

[lxi] Lievens F, Buyse T, Sackett PR. (2005) Retest effects in operational selection settings: Development and test of a framework. *Personnel Psychology;* **58:** 981-1007

[lxii] Raymond, M.R., Swygert, K.A. and Kahraman, N., 2012. Psychometric equivalence of ratings for repeat examinees on a performance assessment for physician licensure. *Journal of Educational Measurement*, **49**:339-361.

[lxiii] McManus IC, Ludka K. (2012) Re-sitting a high-stakes postgraduate medical examination on multiple occasions: nonlinear multilevel modelling of performance in the MRCP (UK) examinations. *BMC Medicine;* **10**:60.

[lxiv] Millman J. (1989) If at first you don't succeed: Setting passing scores when more than one attempt is permitted. *Educational Researcher;* **18**: 5-9

[lxv] Chavez, A.K., Swygert, K.A., Peitzman, S.J. and Raymond, M.R., 2013. Within-Session Score Gains for Repeat Examinees on a Standardized Patient Examination. *Academic Medicine*, **88**:688-692.

[lxvi] Pell G, Fuller R, Homer M, Roberts T. (2012) Is short-term remediation after OSCE failure sustained? A retrospective analysis of the longitudinal attainment of underperforming students in OSCE assessments. *Medical Teacher* **34**: 146-150

[lxvii] Feinberg RA, Raymond MR, Haist SA (2015) Repeat testing effects on credentialing exams: are repeaters misinformed or uninformed? Educational Measurement **34**:34-39.

[lxviii] Tiffin, P.A., Paton, L.W., Mwandigha, L.M., McLachlan, J.C. and Illing, J., (2017). Predicting fitness to practise events in international medical graduates who registered as UK doctors via the Professional and Linguistic Assessments Board (PLAB) system: a national cohort study. *BMC medicine*, **15**:66.

[lxix] Cunningham, S.A., Mitchell, K., Narayan, K.V. and Yusuf, S., (2008). Doctors' strikes and mortality: a review. *Social Science & Medicine* **67**:1784-1788.

[lxx] Hogan H, Healey F, Neale G, Thomson R, Vincent C, Black N. (2012) Preventable deaths due to problems in care in English acute hospitals: a retrospective case record review study. *BMJ quality & safety.* **21**:737–745.

[lxxi] NHS England (2016) Never Event Reports. Accessed 2nd December 2016. https://www.england.nhs.uk/patientsafety/wp-content/uploads/sites/32/2016/01/provsnl-ne-data-2014-15.pdf

[lxxii] GMC (2015) The state of medical education and practice in the UK report: 2015. http://www.gmc-uk.org/Chapter_2_SOMEP_2015.pdf_63501223.pdf

[lxxiii] NCAS (2016) Accessed 2nd December 2016. file://lha-113/pers-H/00078593/Downloads/Report%20to%202013-14%20140530%20website%20(3).pdf

[lxxiv] Millman J. (1989) op. cit.

[lxxv] Clauser BE, Nungster RJ. (2001) Classification accuracy for tests that allow retakes. *Academic Medicine*; **76**: S108-S110

[lxxvi] Millman J. (1989) If at first you don't succeed: Setting passing scores when more than one attempt is permitted. *Educational Researcher;* **18**: 5-9

[lxxvii] Clauser BE, Margolis MJ, Case SM. (2006) Testing for licensure and certification in the professions. In *Educational Measurement* 4 ed. Ed Brennan RL. ACE/Praeger Series on Higher Education

[lxxviii] Tighe J, McManus IC, Dewhurst NG, Chis L, Mucklow J. (2010) The standard error of measurement is a more appropriate measure of quality for postgraduate medical assessments than is reliability: an analysis of MRCP (UK) examinations. *BMC Medical Education*; **10**: 40

[lxxix] Carretta, T. R., Zelenski, W. E., & Ree, M. J. (2000). Basic Attributes Test (BAT) Retest Performance. *Military Psychology,* **12**:221-232.

[lxxx] Reason J (2000) Human errors: models and management. *BMJ* **320**: 768-770.

[lxxxi] Kehoe, A., 2017. A study to explore how interventions support the successful transition of Overseas Medical Graduates to the NHS: Developing and refining theory using realist approaches (Doctoral dissertation, Durham University).