Report on:

SQE Stage One Pilot

30th July 2019

Contents

Report on SQE Stage 1 Pilot

1. Aims

The purpose of Kaplan's work on Stage 1 was to review the SRA's 2017 draft Assessment Specification with a view to making recommendations for a Stage 1 assessment which would be **fair, reliable, accurate, valid, cost effective and manageable.**

In particular the draft 2017 Assessment Specification said answers were needed to the following questions to ensure this overarching aim was met.

What is the minimum number of separate assessments required in Stage 1 to reliably and validly assess the functioning legal knowledge we have set out in the Assessment Specification and Summary of Functioning Legal Knowledge (SFLK)?

What number, format and type of questions and length of assessments will most reliably and validly assess the functioning legal knowledge we have set out in the Assessment Specification and SFLK?

What should the balance of practical legal skills assessments be between Stage 1 and Stage 2? For example, could legal drafting be reliably and validly assessed at Stage 1?

What are the benefits and risks of retaining a non-compensatory standard setting model, as currently proposed, as opposed to a total compensatory or a partial compensatory (within clusters of competence) standard setting model?

In order to investigate these matters Kaplan ran a Stage 1 pilot and at the same time conducted extensive stakeholder consultation on key questions.

2. The Stage 1 pilot

The Stage 1 pilot ran across 3 days from 20 – 22 March 2019. Recruitment of candidates was via the SRA website. Candidates were selected who were broadly representative of those would sit Stage 1 of the SQE both as to prior education and demographic characteristics. 555 candidates were invited to take part in the pilot. 419 accepted their place. 58 of them cancelled in the run up to the examinations and there were 43 no-shows on the day, leaving 318 active participants with 316 sitting all 3 days. 18 candidates requested and were granted reasonable adjustments.

2.1 The Functioning Legal Knowledge (FLK) Test

Candidates sat a total of 360 multiple choice questions which required a single best answer, testing the FLK in the 2017 Draft Assessment Specification. These questions were divided into 3 papers of 120 questions each as follows:

1. Business law and practice; dispute resolution; contract; tort
2. Property practice; wills and the administration of estates and trusts; solicitors accounts; land law
3. Public and administrative law, legal system and legal services, criminal litigation and criminal law

Ethics was examined pervasively throughout the papers.

2.2 The Stage 1 skills test

The skills assessment consisted of a legal research and two writing exercises. For the purposes of the pilot, and to aid decision making, candidates were asked to complete this cycle twice. Results looked at the exam as 2 assessments, referred to as "mini-OSCEs",[1] each containing one research and two writing exercises, also known as stations. In addition and again to aid decision making, all answers were double marked by two independent markers. Unlike the FLK, the passing standard for Stage 1 skills was not that of qualification but "a threshold skill level to enable candidates to work effectively in professional legal services in an unqualified capacity".

3. Stakeholder engagement

Stakeholder views fed into the framing of the key questions mentioned above and the approach taken by the pilot. In particular the key concern that candidates could compensate for lesser knowledge in some areas with higher exam scores in other areas was addressed. In addition Kaplan has sought stakeholder views on key issues: blueprinting (the balance between different subjects); legal content; and the assessment specification. Methods of stakeholder engagement included conferences, surveys, meetings of the reference group and posting to the SQE LinkedIn group. Views on post-pilot recommendations as to assessment design were discussed with the reference group on 3 July.

[1] Objective Structured Clinical Examination (OSCE) is a term frequently used for skills assessments in healthcare and related professions.

4. Results of the pilot and of stakeholder engagement

4.1 FLK

4.1.1 Statistical analysis: descriptive statistics

Across the three 120 question FLK tests candidates marks ranged from 17% to 85% of the total marks available. **Appendix 1** gives the score distributions on each of the three pilot FLK papers.

Average scores were around 50%. It was not considered appropriate to set a pass mark for the pilot because the pilot was not aiming to test whether candidates passed or failed. The pass mark for the SQE will vary between exams, to ensure that the standard of the assessment remains consistent from one sitting to the next. However, on these pilot questions the pass mark would likely have been above 50%.

4.1.2 Statistical analysis: reliability, accuracy and cost-effectiveness

Reliability and accuracy of outcomes are key quality measures of a high stakes licensing exam. Reliability is about how reproducible the results would be on another test. All measurements contain some error, and accuracy is about how large this error is. A regulator needs to be sure that the candidates who pass deserve to pass, and those who fail deserve to fail.

Analyses, including generalisability analysis, show scores of three exams of 120 questions were slightly less reliable and accurate than the level commonly regarded as the "gold standard" in national licensing exams[2]. In addition, it should be noted that estimates derived from a pilot might be lower in the real exam because the range of scores is likely to be reduced in a live setting. Candidates will do more preparation eliminating some of the low scores.

For analysis purposes therefore, we also divided questions into two exams of 180 questions each as follows:

1. Business law and practice; dispute resolution; contract; tort, public and administrative law, legal system and legal services.
2. Property practice; wills and the administration of estates and trusts; solicitors accounts; land law, criminal litigation and criminal law .

Scores based on two exams of 180 items each provided better statistics and did reach the levels of reliability and accuracy commonly considered desirable in national licensing exams.

[2] For example "..there is a consensus among medical educationalists that high stakes assessments, such as most of the Royal College examinations, should have a reliability of at least 0.9." Page 36, Postgraduate Medical Education and Training Board: *Developing and maintaining an assessment system - a PMETB guide to good practice* London:PMETB; 2007

A key concern of stakeholders is the extent to which examinees could compensate for lesser knowledge in some areas with higher exam scores in other areas. Statistical analyses showed that good candidates tend to do well in all sections, and weak candidates tend to do badly. On tests of both 120 and 180 questions, compensation by performing well enough in some areas to offset other areas can happen, but is uncommon, and the scores compensated for tend to be marginal. Compensation was very slightly more marked in 2 exams of 180 questions than in 3 exams of 120 questions but not enough to cause concern. However the extent of compensation is of importance to stakeholders and should be reviewed in all live deliveries of the SQE to ensure it remains uncommon and that the scores compensated for are marginal. **Appendix 2** gives an analysis of compensation looking at tests of both 120 and 180 questions.

Cost effectiveness will be achieved with the minimum length of assessment to reliably and accurately assess the FLK. Test design is always a compromise between competing factors of which cost, reliability, precision, number of questions, and the extent of compensation tolerated between subjects are key. The foregoing investigations give us the evidence to make a balanced decision.

4.1.3 Further validity evidence

Validity of the questions in the pilot was further established through candidate feedback, review by the Kaplan academic team of solicitors, and by the Angoff Panel (9 solicitors drawn from practice and teaching who reviewed each question for the purpose of advising on the pass mark).

The SRA conducted a stakeholder survey on blueprinting of the FLK. Opinions on the correct weighting of subjects differed widely. For instance the weighting suggested for property varied between 10% and 50%; that for business law between 10% and 40%, that for contract between 10% and 40%. This was one factor taken into account in determining the blueprint.

There has also been extensive consultation on the FLK with experienced QLTS assessment writers, delegates at the SQE conference in December 2018, the SQE Reference group, and a survey posted to the SQE LinkedIn Group. This work was not completed in time for the Stage 1 pilot but will feed into the revised Assessment Specification.

4.1.4 Fairness: demographic and equality analysis

Kaplan undertook a series of statistical analyses on these issues. These analyses should be viewed with caution given the small numbers of candidates involved, confounding variables, (the fact that some categories had a very significant overlap with other categories), the fact that characteristics were self-declared, the fact that behaviour in a pilot will be different from that in a live exam, and for some analyses (particularly that of sources of score variance), the complexity of statistical modelling.

Appendix 3 shows the marks achieved by candidates by ethnicity, gender and disability on the pilot FLK. We would comment as follows:

Ethnicity: There were differences in performance by ethnicity in the pilot FLK with BAME candidates performing worse overall than white candidates. Exploratory statistical analyses were conducted to try to determine the sources of score variance between candidates. There are many overlapping variables which may be significant. In the pilot FLK this analysis suggested that factors such as completion of the GDL and completion of a law degree at a Russell Group University were a more significant source of score variance than ethnicity. Kaplan and the SRA will continually monitor any disparity in performance between protected groups and will continue to review questions before and after use with a view to promoting equality of opportunity.

Gender: There were also differences in performance by gender in the pilot FLK with women overall performing slightly worse as a group compared to men. These differences were not very marked but varied in extent by paper and throughout the score distribution. Exploratory analysis of sources of score variance in the pilot FLK suggested that gender was of limited, and only limited, significance. As explained above it suggested that the most significant sources of score variance were completion of the GDL and completion of a law degree at a Russell Group university. Data from QLTS provide a point of reference here. They fluctuate year on year as to whether men or women do better in the equivalent of the FLK. Again, Kaplan and the SRA will continue to monitor these issues with a view to promoting equality including review of questions before and after use.

Disability: there were also differences in performance by disability in the pilot FLK. However, numbers who declared themselves disabled under the Equality Act 2010 were too few for any conclusions to be drawn. Numbers will be larger following implementation when these analyses will be performed again to monitor any differential between groups in observed performance standards.

4.1.5 Manageability

The FLK pilot was run successfully at 44 Pearson VUE test centres spread throughout the UK and in Singapore and France. There were no major incidents. One room in one test centre in London was unusable on one day. In line with the Business Continuity and Incident Response Plan the incident was quickly resolved with the small number of candidates involved being relocated with minimal disruption. A test screen resolution issue also arose for some candidates and for the majority the resolution was either corrected or candidates were moved to other computers. Prior testing will be conducted and clearer instructions provided to test centres regarding resolution in future. A variety of reasonable adjustments were accommodated including extra time and testing in individual rooms. In summary, running SQE Stage 1 FLK exams using both domestic and international Pearson VUE test centres is both feasible and manageable including making appropriate reasonable adjustments for candidates where required. A comprehensive reasonable adjustments policy will be further developed before implementation based on relevant good practice guidelines. Further systems also need to be developed to deliver the FLK assessments efficiently at scale but the Pearson VUE test centre network has adequate capacity and the period before implementation in 2021 allows ample time to develop, test and document these systems.

4.2 Stage 1 Skills

4.2.1 Statistical analysis: Descriptive Statistics

Candidate marks for each mini-OSCE (see above) ranged from 8% to 100%. However, performance in the two "mini OSCEs" was very different with mean scores of 75.8% and 59.1% respectively. **Appendix 4** shows the contrast between distributions of candidate scores on the two pilot "mini-OSCEs"

This disparity between deliveries would be challenging to defend in a high stakes professional exam.

Further analysis showed that candidate scores on one station were much lower than on others. It was necessary to examine whether candidates' low marks reflected a poorly drafted question, poor marking, or lack of knowledge. Two independent markers marked the station almost identically and raised no issues about the question. This suggested strongly that the results were not due to a poorly drafted question or poor marking. This view was supported by our own review of the station's content and expected standard. The question concerned taxation issues, and as candidate performance on the FLK was also poor on taxation issues it appears that poor performance related to tax persisted across assessment formats.

More importantly for the pilot, this draws attention to a major issue with a pass/fail decision based on so few stations. One station performing differently can have an unacceptable effect on outcomes. The same would apply if for instance there were deficiencies in the question, or problems with delivery. With so few stations, it would be difficult to defend withdrawing the assessment from the final mark. The design of a high stakes professional exam should ensure it is sufficiently robust and resilient to withstand unusual performance in a station for whatever reason.

4.2.2 Statistical analysis: reliability and accuracy

As explained above, reliability (reproducibility) and accuracy of outcomes (precision) are key to a high stakes licensing exam. Regulators must ensure that the candidates who pass deserve to pass and candidates who fail deserve to fail. The mini OSCEs were remarkably reliable for an exam with so few stations but they did not reach the standards of reliability commonly regarded as necessary for a national licensing exam and fell far short of those for accuracy.

4.2.3 Fairness: demographic and equality Issues

Appendix 5 shows the marks achieved by candidates by ethnicity, gender and disability on the pilot Stage 1 skills. We would comment as follows:

Again analyses should be viewed with caution because of the small numbers of candidates involved, confounding variables, (the fact that some categories had a very significant overlap with other categories), the fact that characteristics were self-declared, the fact that behaviour in a pilot will be different from that in a live exam, and for some analyses (particularly that of sources of score variance), the complexity of statistical modelling.

Ethnicity: There were differences in performance by ethnicity in the pilot Stage 1 skills with BAME candidates performing worse as a group overall compared to white candidates. Exploratory statistical analyses were conducted to try to determine the sources of score variance between candidates. They suggested that ethnicity was the most significant source of score variance for both Mini OSCEs in the Stage 1 pilot skills.

We therefore looked at the performance of white candidates compared to BAME candidates taking into account their MCT performance. This analysis suggested that the pilot Stage 1 skills advantaged those of white over BAME ethnicity even taking into account MCT scores.

The various indications that the Stage 1 skills produce a discriminatory effect between white and BAME candidates suggest that advancing equality of opportunity calls into question whether a Stage 1 skills assessment in its present form can be justified.

Gender: There was little difference in performance by gender in the Stage 1 pilot skills, with no significant contribution to score variance in either mini-OSCE

Disability: There were differences in performance by disability in the pilot Stage 1 skills. However, numbers who declared themselves disabled under the Equality Act 2010 were too few for any conclusions to be drawn. Numbers will be larger following implementation when these analyses will be performed again.

4.2.4 Validity: setting the passing standard

Setting a valid passing standard for Stage 1 skills was also problematic. The passing standard was to be a “threshold skill level to enable candidates to work effectively in professional legal services in an unqualified capacity”. This however raised considerable conceptual difficulty as to exactly what the standard was. An “unqualified capacity” can have a wide variety of meanings. It was also unclear what place a pass/fail exam at a different level had within a licensing exam.

5. Recommendations and next steps

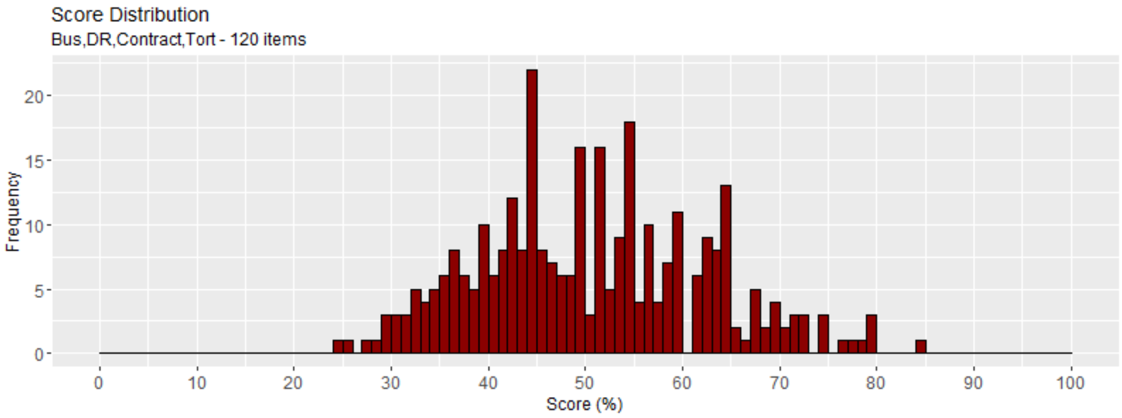
For the foregoing reasons we make the following recommendations for **a fair, valid, reliable, accurate, cost effective and manageable** Stage 1 SQE exam**:**

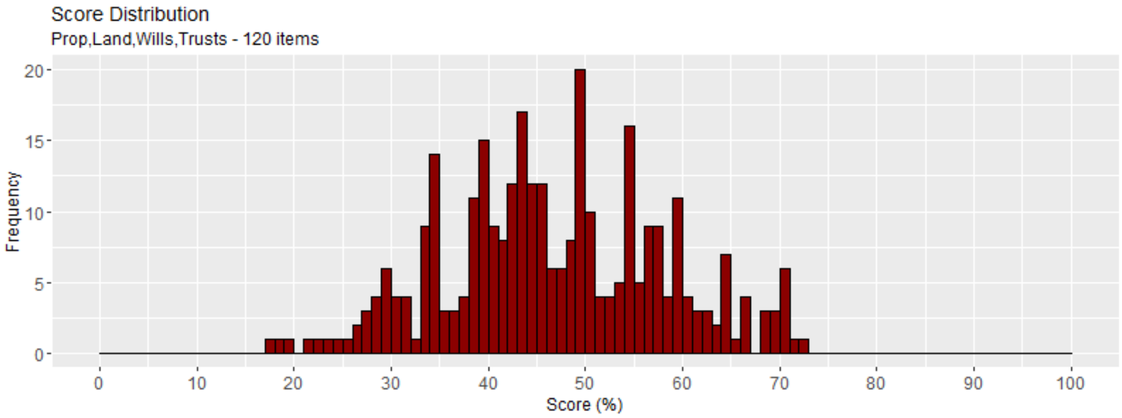
2 x 180 question FLK multiple choice tests requiring a single best answer. This recommendation is borne out by the statistical analysis of the pilot and the analysis of compensation. However, some stakeholders feel more comfortable with 3 x 120 question FLK and such a design would be acceptable though not optimal.

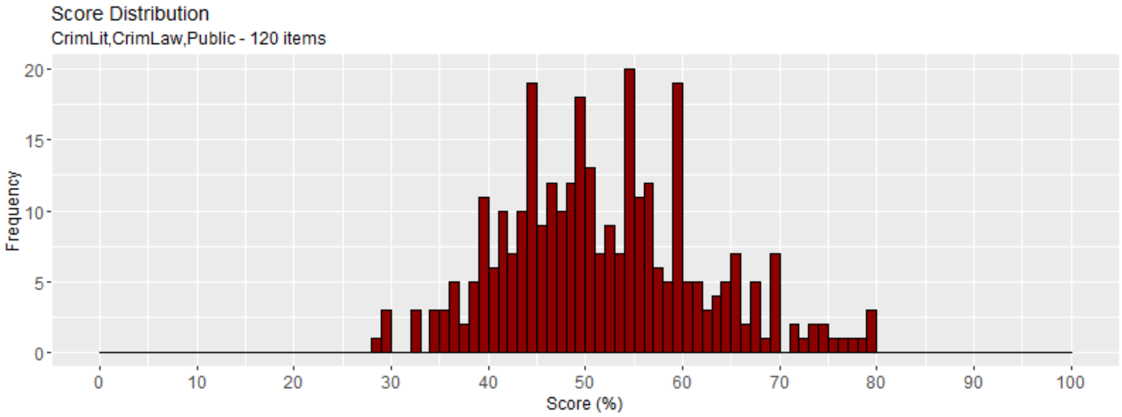
For reasons of the robustness and resilience of the exam as well as reliability, accuracy, validity, fairness and equality of opportunity, a Stage 1 skills exam in its current form should not be part of the SQE.

The next steps will be further discussion with stakeholders about Stage 1 skills and possible alternatives to the current model, and publication of a revised Stage 1 Assessment Specification.

Appendix 1: Candidates’ scores on the Stage 1 pilot FLK







Appendix 2: Subject score quintiles and overall FLK test score

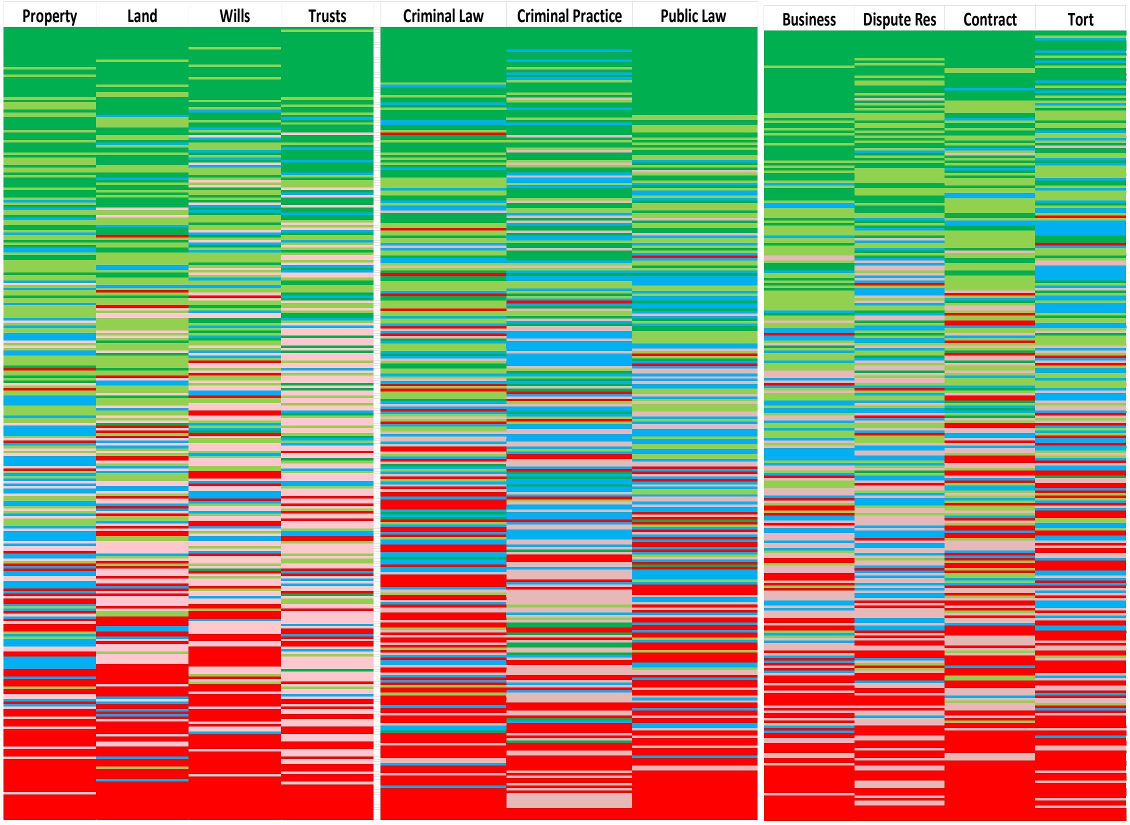
The following figures provide a visual description of the subtest performance of candidates on the pilot FLK exam. This performance is illustrated both for the exam considered as 3 x 120 question tests, and for it considered as 2 x 180 question tests.

The figures are organised vertically by ascending total score. In addition to the total score, performance on the sub-test components is shown on the basis of the candidates being divided into quintiles on their performance on the sub-test.

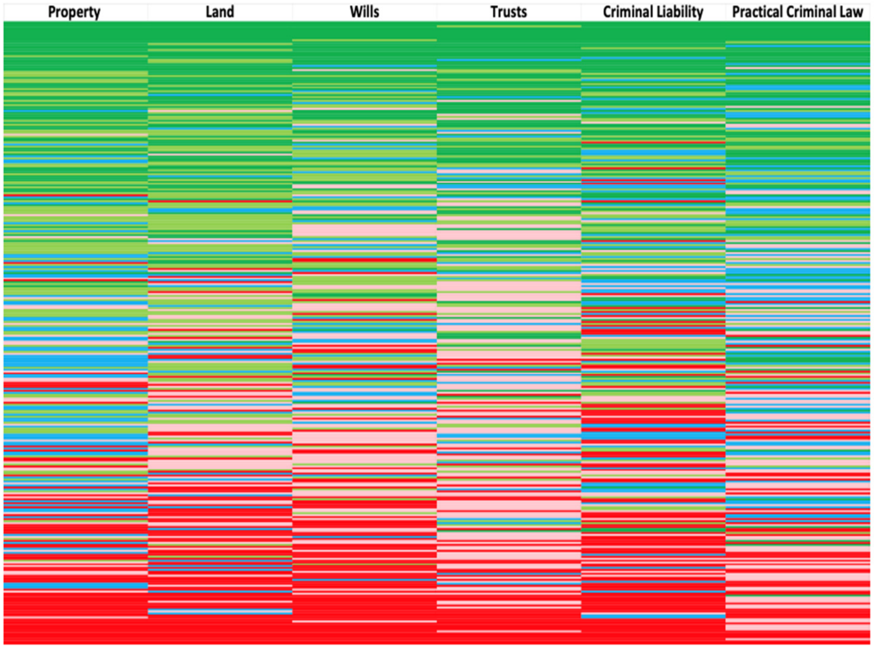
Those with the top scores were coloured green, those in the next quintile were coloured blue, next yellow, next pink, and those in the bottom quintile were coloured red. If candidates were in the same quintile across all subjects, the figure would have a solid green band; below that would be a solid blue band; etc, and the bottom band would be solid red. As expected, the figures show that some candidates had higher scores on some topics than others; there are some yellow bands in the otherwise green area for example.

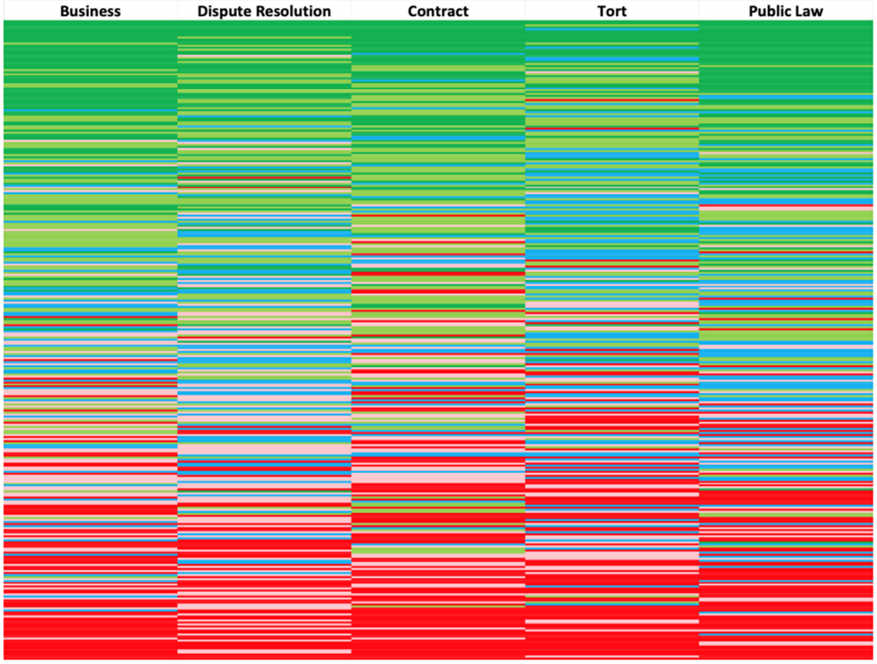
However, it also shows that the extent of compensation between subject areas is limited. Good candidates tend to do well overall and bad candidates badly. It is plausible that in a real exam, scores would be even more uniform (although not entirely uniform) across subjects as candidates will have spent more time preparing for the exam and rely less on their existing knowledge.

3 x 120 question tests

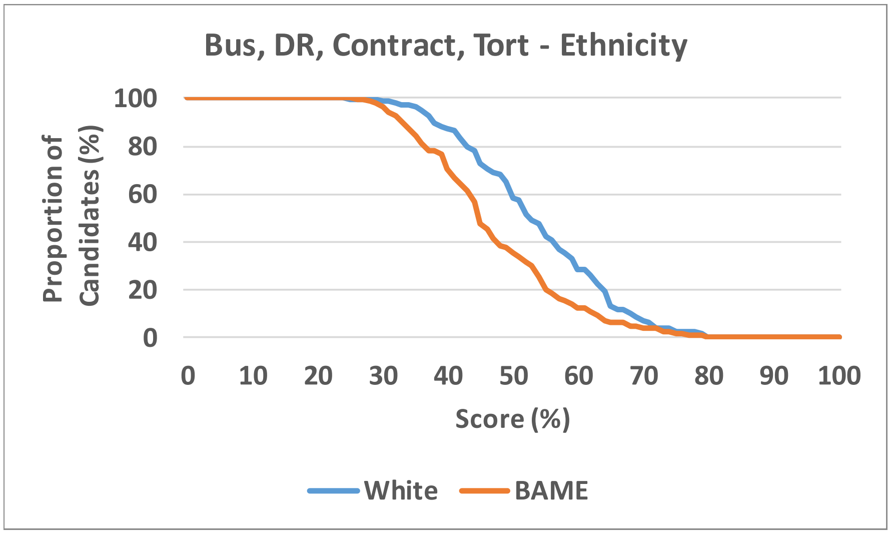


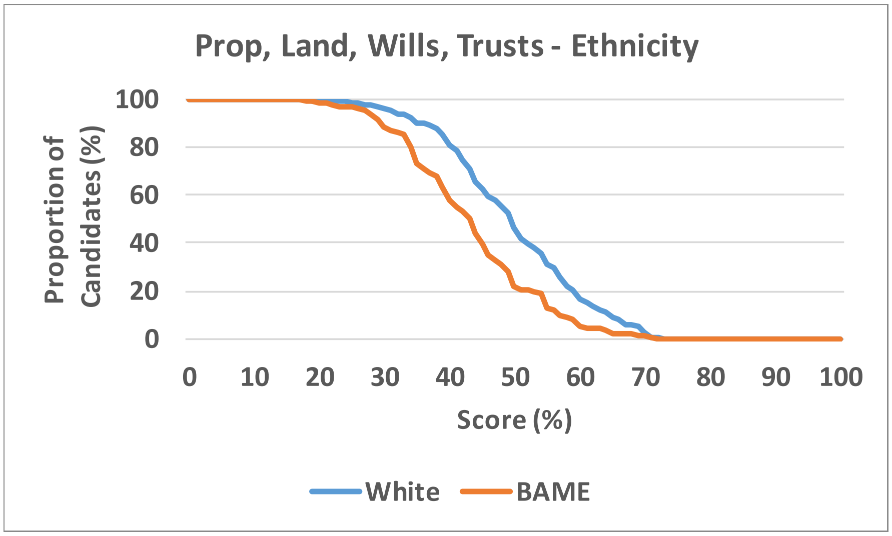
2 x 180 question tests

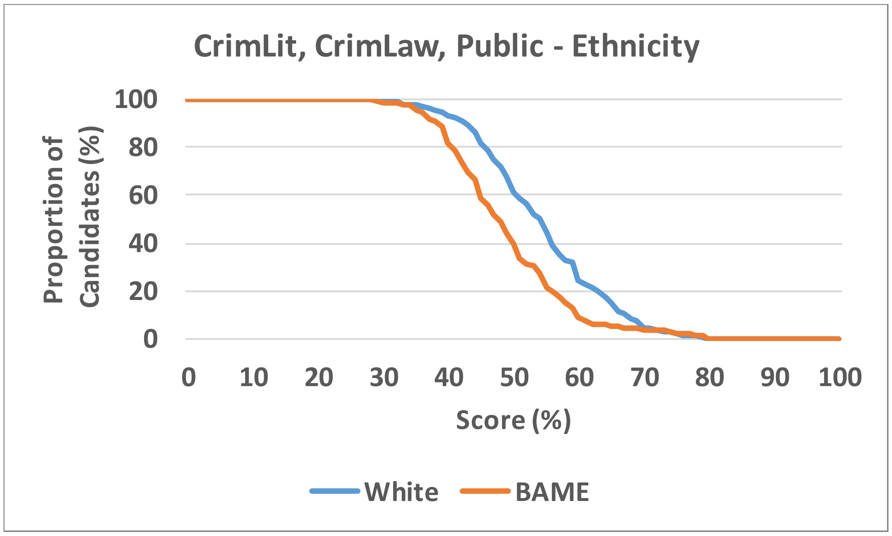


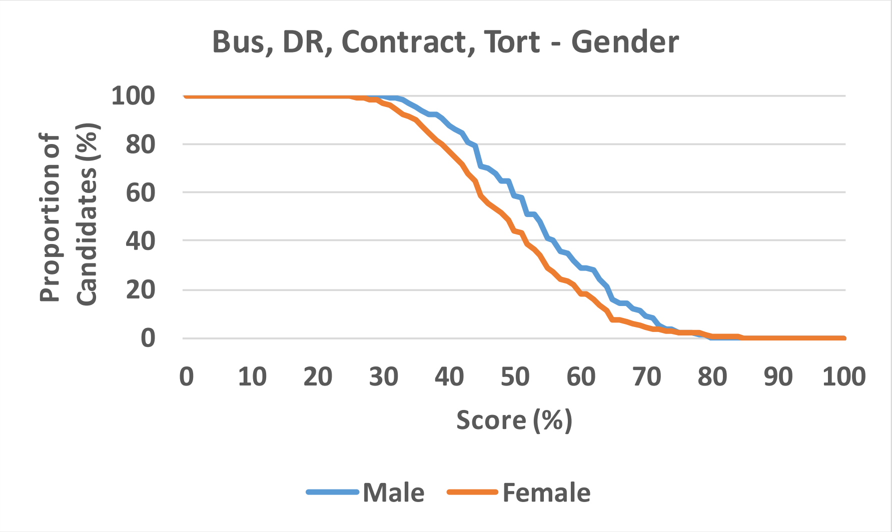


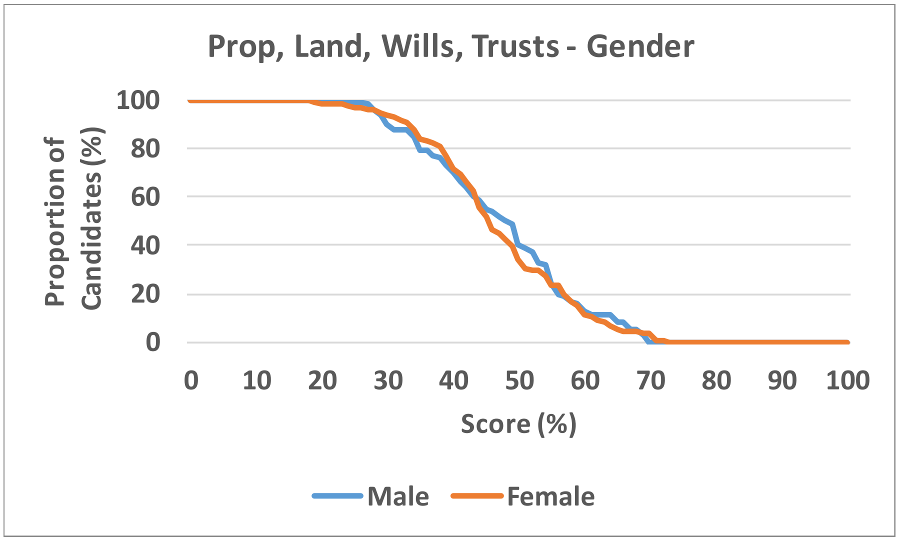
Appendix 3: Marks in the pilot FLK by ethnicity, gender and disability. (The small numbers, especially of candidates declaring a disability, mean these should be viewed with caution.)

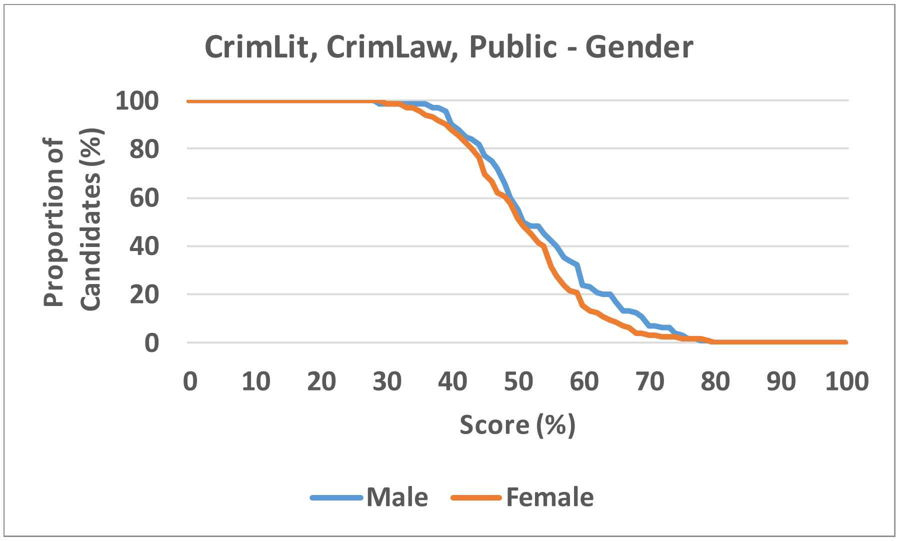


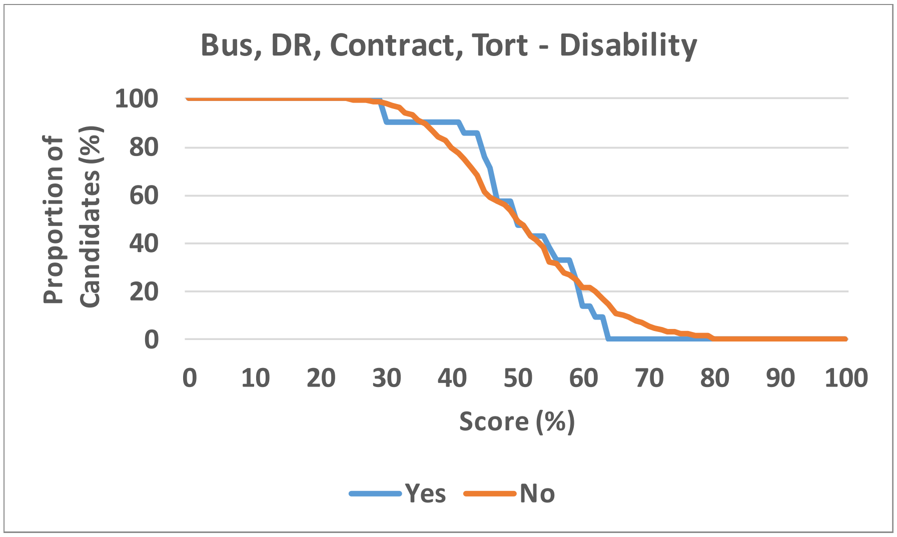


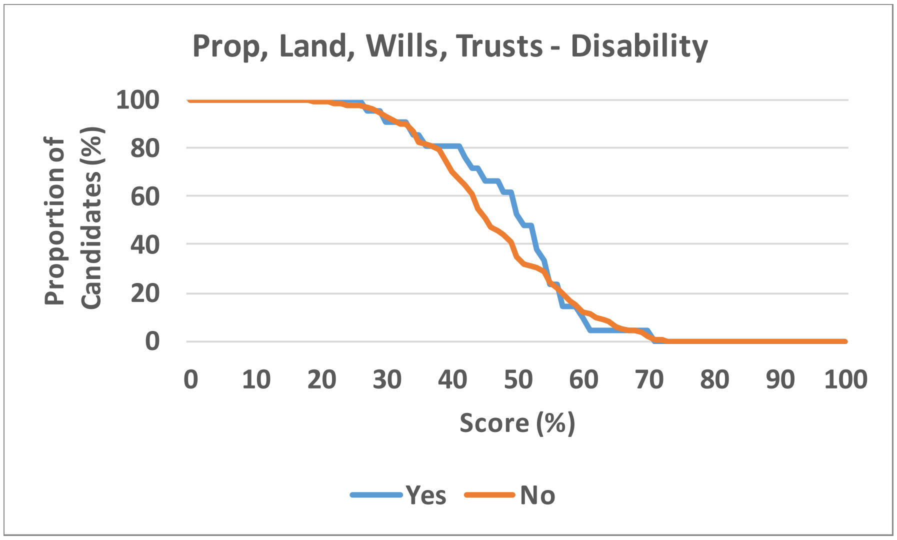


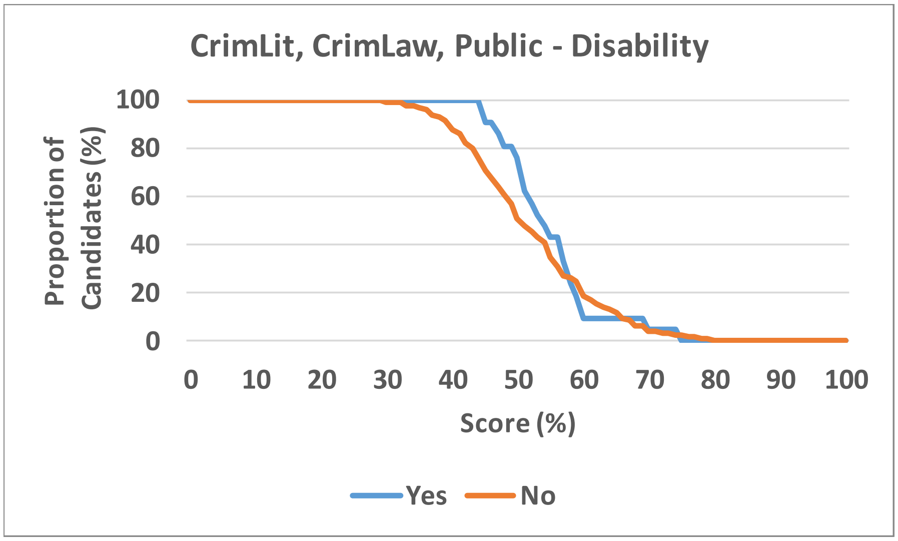

Bus, DR, Contract, Tort - Gender
Proportion of Candidates (%)
100
80
60
40
20
0
0
10
20
30
40
50
60
70
80
90
100
Score (%)
Male
Female


Prop, Land, Wills, Trusts - Gender
Proportion of Candidates (%)
100
80
60
40
20
0
0
10
20
30
40
50
60
70
80
90
100
Score (%)
Male
Female

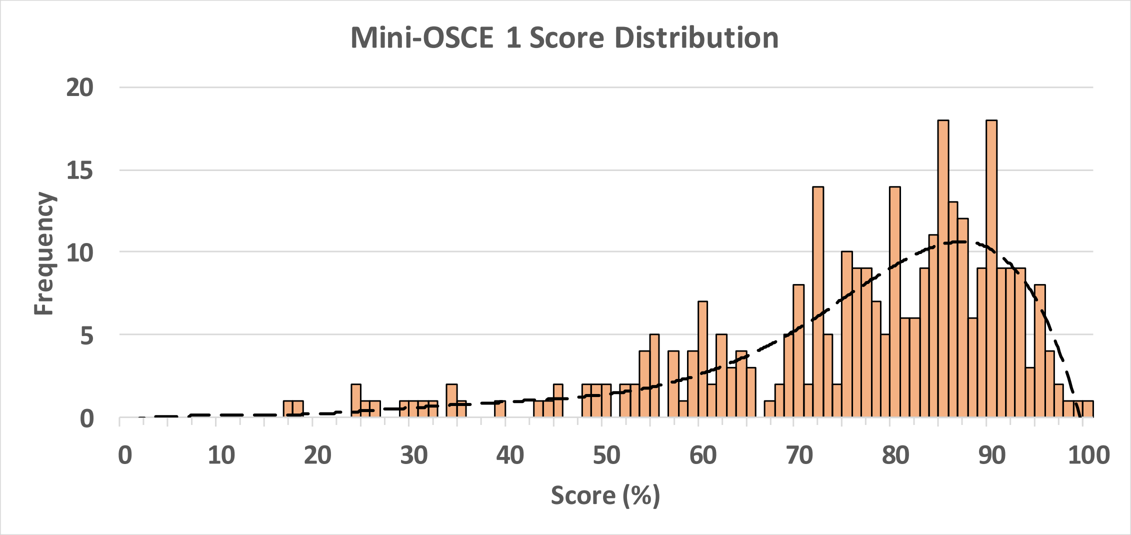

CrimLit, CrimLaw, Public - Gender
Proportion of Candidates (%)
100
80
60
40
20
0
0
10
20
30
40
50
60
70
80
90
100
Score (%)
Male
Female

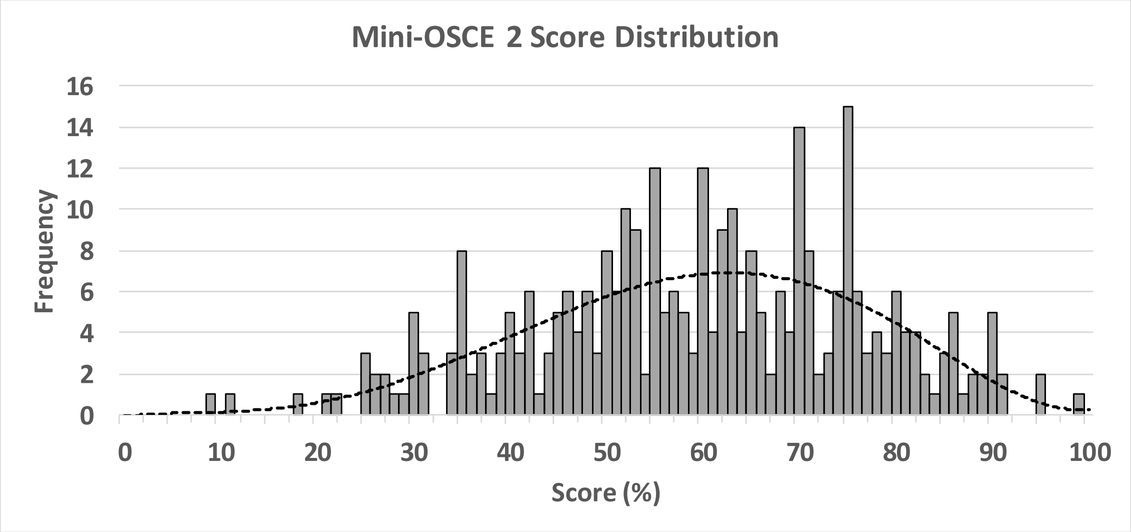

Bus, DR, Contract, Tort - Disability
Proportion of Candidates (%)
100
80
60
40
20
0
0
10
20
30
40
50
60
70
80
90
100
Score (%)
Yes
No


Prop, Land, Wills, Trusts - Disability
Proportion of Candidates (%)
100
80
60
40
20
0
0
10
20
30
40
50
60
70
80
90
100
Score (%)
Yes
No

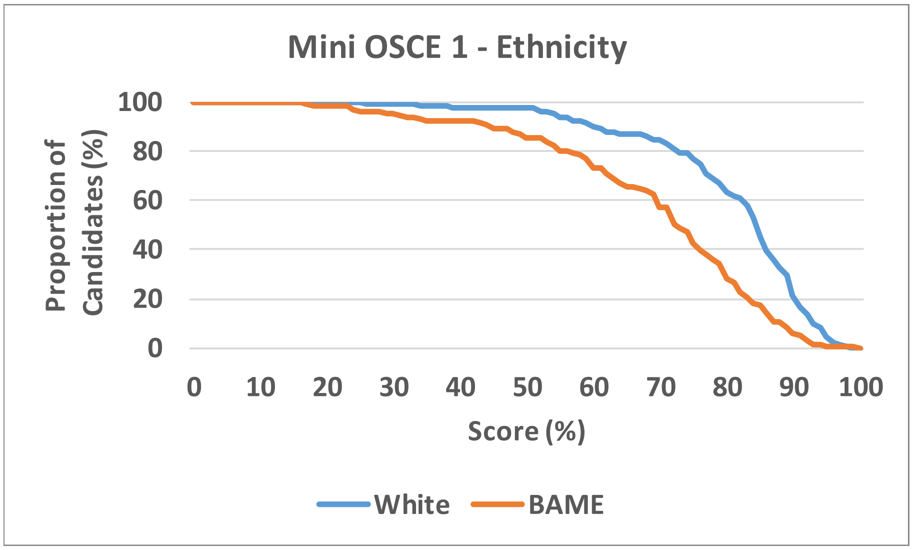

CrimLit, CrimLaw, Public - Disability
Proportion of Candidates (%)
100
80
60
40
20
0
0
10
20
30
40
50
60
70
80
90
100
Score (%)
Yes
No

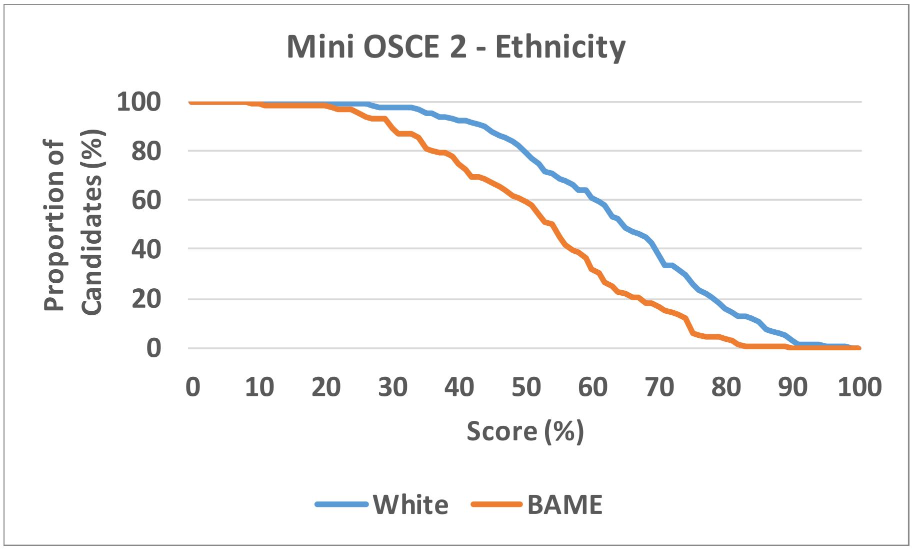
Appendix 4: Candidates' scores on the two pilot "mini-OSCEs"

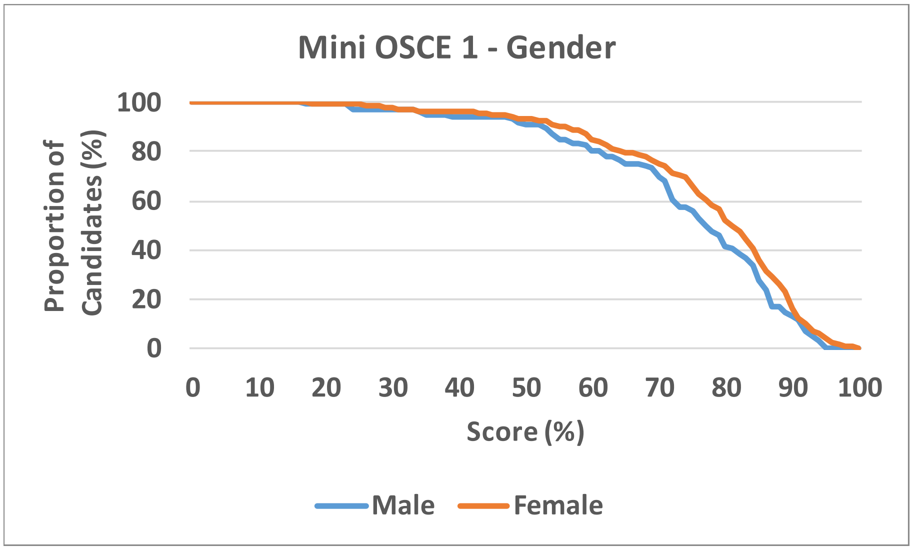


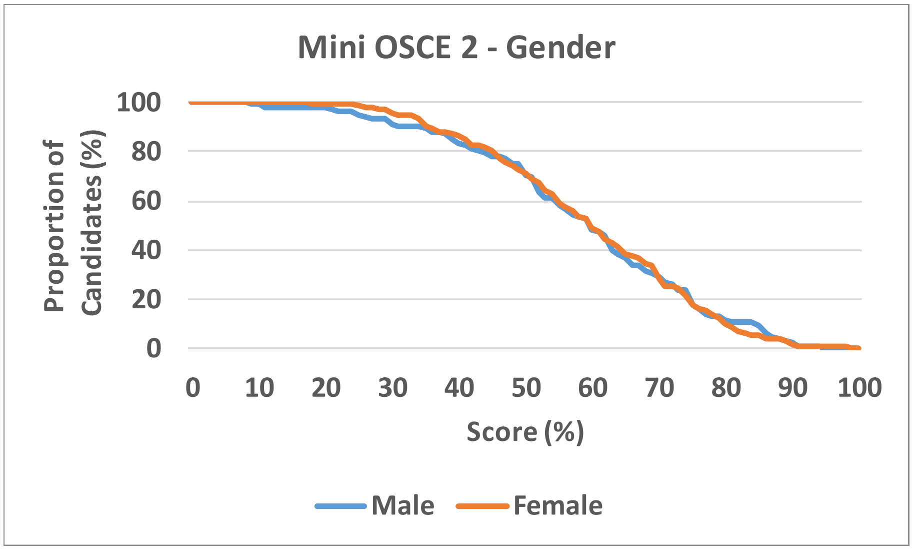


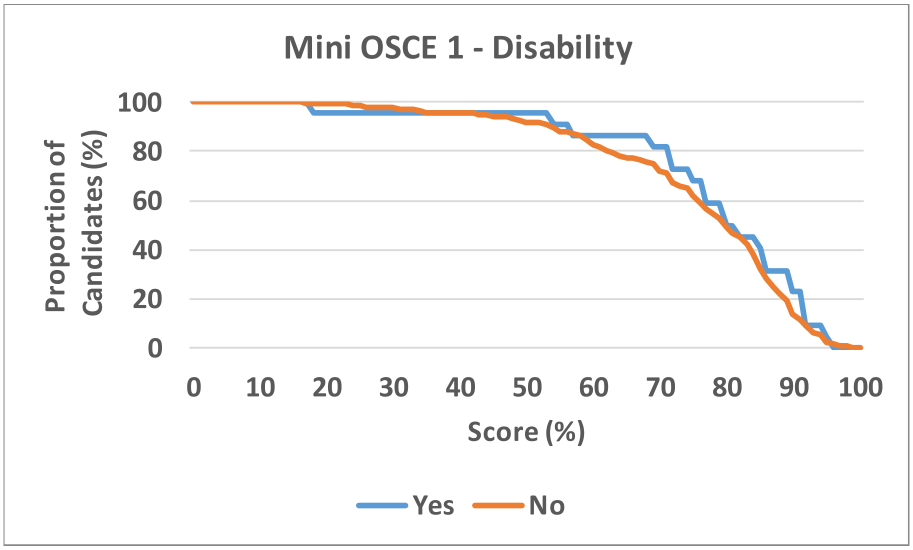
Appendix 5: Marks in the pilot Stage 1 skills by ethnicity, gender and disability. (The small numbers, especially of candidates declaring a disability, mean these should be viewed with caution.)

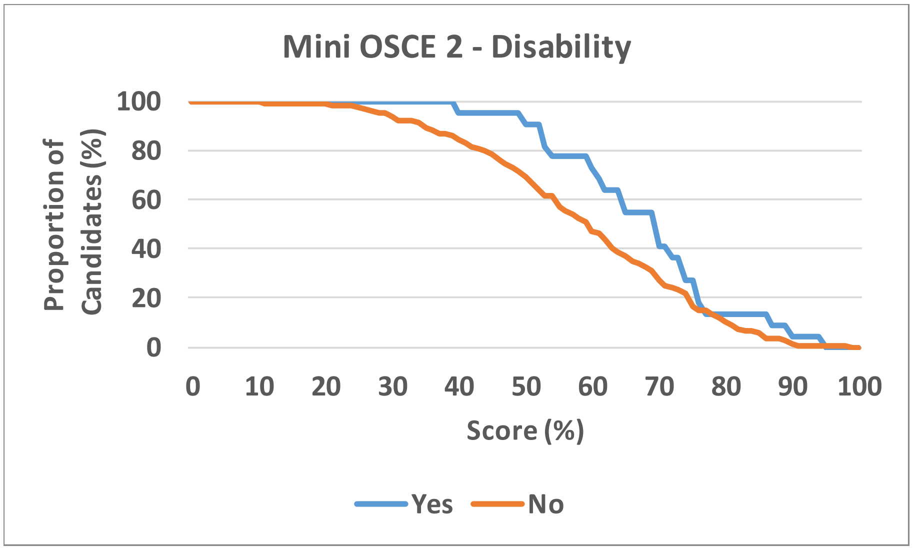





Mini OSCE 1 - Gender
Proportion of Candidates (%)
100
80
60
40
20
0
0
10
20
30
40
50
60
70
80
90
100
Score (%)
Male
Female


Mini OSCE 2 - Gender
Proportion of Candidates (%)
100
80
60
40
20
0
0
10
20
30
40
50
60
70
80
90
100
Score (%)
Male
Female


Mini OSCE 1 - Disability
Proportion of Candidates (%)
100
80
60
40
20
0
0
10
20
30
40
50
60
70
80
90
100
Score (%)
Yes
No


Mini OSCE 2 - Disability
Proportion of Candidates (%)
100
80
60
40
20
0
0
10
20
30
40
50
60
70
80
90
100
Score (%)
Yes
No