SQE2 Pilot

Overview and recommendations from the Independent Reviewer

Final - May 2020

1 Executive Summary

The SQE2 pilot set out to test the model of assessment to be used for the live SQE2. The outcomes of the pilot helpfully contributed to the evidence needed to finalise the SQE2 design, even if the numbers of candidates that participated fully were lower than ideal. However, complementary evidence, which supplements and supports the pilot evidence, enables a design recommendation for SQE2 to be determined.

The planning, operation and analyses of the pilot, were generally of a high or very high quality. All stakeholders want this to be a world class professional qualifying examination. The recommendations made here should be read in a spirit of achieving that, rightfully, high aspiration.

The main conclusion from all the evidence available, not just the indicative evidence from the SQE2 pilot, is the most reliable, valid, fair and defensible model for SQE2 to adopt is the uniform model. In this model candidates take exactly the same assessments or stations, which sample across the relevant key skills and different legal contexts. If this qualification is to achieve its aim of raising the bar of, or at least levelling up, the quality and standards required for solicitor qualification, the universal model provides the safest way forward, particularly with regard to standard setting. This is also a model adopted by most, if not all, high stakes, professional qualifications leading to licensure.

This means that in planning for the live SQE, taking in to account the prior learning from the SQE1 pilot, a range of design options have been tested in these pilot phases. This has enabled Kaplan and the SRA to use this evidence, and evidence from other professional exam contexts, to determine the final design. I have done the same, and a design which has the following features appears to offer the best opportunity for a fair, reliable, valid and defensible examination:

- SQE1 to be 2 x 180 item functioning legal knowledge single response answer tests, with no skills assessment
- SQE2 to be a uniform exam, comprised of between 15 to 18 assessment stations to test relevant legal skills expected of a day one solicitor

This design also offers the best opportunity to be manageable over time, for example in avoiding tasks becoming predictable, while creating a rigorous assessment of each candidate. As with any new qualification the assessment design should be carefully reviewed after each live sitting to analyse where improvement can be made, while ensuring fairness for the first and every subsequent cohort to take it.

2 Preparation for the pilot

2.1 SQE2 Pilot assessment specification

The design of the SQE2 was carefully considered. In the lead up to the pilot there were differing views on the optimal design of SQE2, with influential stakeholders expressing strong, but differing, opinions. The debate was whether SQE2 should be a uniform design, where all candidates take exactly the same assessments sampling across the legal specialisms, or a common core which all candidates take and then choose their preferred legal specialism or complete candidate choice of two legal specialisms out of the five available. The final design of the pilot settled on a core with candidates able to choose one of two specialisms – criminal and business. This enabled learning from doing and the performance data generated helps to inform the final SQE2 design decision. The final assessment specification for the SQE2 pilot was published and made available to pilot candidates approximately 6 weeks before the exams took place.

2.2 Recruitment of pilot candidates

Significant efforts were made to recruit as many pilot candidates as possible for the SQE2. The recruitment process was more challenging than the SQE1 pilot because the expectations and time commitments of candidates were very demanding. Ultimately a lot of candidates that signed up dropped out before and during the pilot. In the end 167 candidates completed all parts of the SQE2 pilot, while greater numbers sitting would have been better, this number does provide overall indicative performance data which is valuable when finalising the SQE2 design.

2.3 Operational readiness

Several weeks before the pilot took place, I met with a range of Kaplan staff who had responsibility for the SQE2 pilot operation. I was provided with a wide range of materials which were being used to support operational readiness and/or explained the rationale to the approach being planned. Overall operational planning was to a high standard, helped by some prior experience of activities that Kaplan have in broadly similar Qualifying Lawyer Transfer Scheme (QLTS) settings.

3 Training and standardisation of assessors

I observed two of the three days of assessor training and standardisation activities at Kaplan's offices in the week leading up to the SQE2 pilot exams being sat.

On 11 December 2019 I observed the training of actor assessors who would be role playing one of the two client interviewing tasks (or stations). 31 actors were used, all were formally trained and Equity card holders and all were sourced from the same agency. The agency had previously been used by Kaplan for QLTS assessments. The session was led by an academic head at Kaplan who did an excellent job of setting the scene, ensuring the assessors understood what was expected of them. Exactly the right balance was struck between progressing through the activities planned while providing time for high quality discussion about the marking standards to be adopted.

Each actor observed the role play between an actor assessor and a solicitor, usually the author of the assessment, who fulfilled the role of the candidate. After observing the role play, each assessor independently determined their domain score for the performance they observed using the common marking criteria. These scores were captured and displayed so all participants could see what scores were being given. Once all 31 actors had provided their assessment scores the academic head led a good discussion by asking the outliers (those who had provided a higher or lower score than others) to explain their reasons. Through this process a very good discussion about where the 'true' mark should lie was facilitated. This, ultimately, is the purpose of such standardisation activities as there was evidence of convergence of marking judgements across the group towards the end of the session.

I observed other excellent practice during this session:

- Explanation of how the assessor should use the incident log should anything unusual happen during the interview with the candidate

- Actors had been carefully selected to provide both a broad representation of ethnicity while being accurate to the role and scenario they were acting, eg all looked of a similar age as the role required
- Helpful explanations about how to conduct the interview in as similar a way as possible regardless of the actor or candidate involved, e.g. whether and when to interrupt the candidate
- In discussing how to determine their score, the actor assessors were told not to be overly influenced by one 'trigger point', for example a closed question or an interruption to instantly inform judgement, and how the global scoring for each candidate should work was explained
- Two facilitators from the agency participated in the training day. They concentrated on ensuring every actor was as similar in their acting performance as possible, offering observation comments about how to present themselves consistently and objectively eg in using emotions
- There was evidence that, as would be expected of professional actors, all had learned their role and scripts and had paid attention to the detailed preparatory notes they had been given
- A buddy system whereby new actors were teamed up with more experienced actors, who had completed assessments in the QLTS or other relevant settings, to offer support and advice

I therefore recommend that these high quality activities are sustained in the live SQE2.

Three role plays were assessed, each enabled a wide range of scores to be observed for each of the different elements of the marking criteria. However, none were assessed as 'fail' or 'marginal fail' for the global standard. **I therefore recommend for the live SQE2 a wider spread of overall performance on the global marking scale is planned for future standardisation and training, recognising that 'marginal fail' and marginal 'pass' exemplars are always important to include.**

I also recommend Kaplan video record the standardisation role plays and make these videos available to assessors after the standardisation and training days. This will offer assessors a reference source to exemplify marking standards. Actors could refer to these videos prior to assessments taking place, and during any downtime, such as the evening before assessments commence, to help them maintain consistent standards throughout their assessment windows. Such evidence would also provide a usual archive to maintain standards over time.

While there was some discussion about the importance of treating all candidates the same in the training, I did not observe any advice offered to the actors about the risks of unconscious bias in their assessment responsibilities. I questioned this on the day and Kaplan have subsequently shared with me their high level plans to offer training and support on how to avoid unconscious bias in the future, amongst a comprehensive range of other measures, to reduce the risk of any unfairness to any candidate as a result of their ethnicity, religion or background. **It is important these plans are executed.**

On 12 December 2019 I observed the training of solicitor assessors who were responsible for the two advocacy/oral presentation tasks/stations. The format followed a similar structure to the previous day for the actor assessors. 29 solicitors were trained and once again role plays were conducted and each assessor independently scored each element of the marking criteria across three assessments, in business, property and criminal. There was a good discussion about the extent to which 'correct application of the law' should influence marking and about how important it was to review the overall pass mark (the global score). For example, the academic head leading the session made it clear that evidence of very good skills could compensate for imprecise application of the law when determining the global score. It is important that a clear distinction is drawn between capturing the domain scores and the global scores, to avoid there being compensation between the two because this would compromise the borderline regression process**. I recommend Kaplan ensure there is always an explanation to assessors, especially any new assessors, about how the pass mark is set and the impact of their actions in the live context.**

Given each solicitor tends to have expertise in certain contexts of law, there is a risk their deep knowledge in certain areas could lead them to inconsistent assessment judgements. This may be due to an emotional reaction to what they know to be right or wrong in the precise application of law in their specialist area, which another assessor who is not a specialist, might not have. **I therefore recommend the narrative advice, which sits alongside the published marking criteria descriptions, about how to form judgements in the live SQE2 fully clarifies how to interpret the 'application of the law' marking criteria when making domain scoring judgements, which is especially important for any new assessors.**

A fine balance needs to be struck between simulating what is reasonable to expect a Day 1 solicitor to know about application of the law, what additional information is offered to them on the day as part of the assessment and how to assess each skills based performance by the candidate. Designing the assessment task to be authentic to that likely to be faced by a Day 1 solicitor involves very careful preparation. The narrative advice to solicitor assessors will need to make clear if any basic lack of knowledge of application of the law eg that would be dangerous to advise a client, or advice that was unethical, should not be allowed to pass when providing the global score for this task.

During this training session some solicitors expressed concern about the nature of the provision of some of the materials made available on the day for pilot candidates to prepare for the task. Specifically, in one task candidates were provided with parts of the Companies Act, it was noted that other parts of the Act (not provided to candidates) made earlier paragraphs voidable. It was therefore considered a risk to pick out certain elements of law in case this misled candidates. It should be noted that late changes were made to at least one of the SQE2 pilot tasks to include more supporting materials. It is well documented in the academic literature that making late changes to assessments in this way heightens the risk of introducing error and/or construct irrelevant variance. **I therefore recommend that the provision of materials to support assessment tasks is very carefully considered to reduce risks. There is a need to strike the right balance between ensuring this is a skills-based assessment, without candidates being expected to remember lots of detailed legal knowledge, while expecting basic legal concepts and application of the law to be understood and applied. I therefore recommend the advice made available to candidates preparing for the SQE2 is clear about what are the legal concepts and knowledge they are expected to know and all options for reducing the preparatory burden for candidates are considered. It is entirely appropriate that some tasks do, and some do not, have supporting materials, according to the objectives of that assessment task and its construct.**

4 Observation of assessments being completed by pilot candidates

On 13 December 2019 I observed the assessment of candidates taking place. The scale of the task is significant because of the complexity of having the right assessors, the right candidates and the appropriate staff facilitating the whole day in the right place at the right time. It was organised with military precision and was a highly impressive feat of organisation.

I observed:

- a briefing being given to candidates
- the role play assessments being conducted, this was remotely in the video observation room
- briefing to candidates with individual candidate exceptional requirements (ICER)
- the feedback session with actors after they had completed their initial assessments

The venue had an observation room where academic heads and actor facilitators could monitor and observe in real time the assessments taking place for any candidate. This enabled the smallest detail eg how the room is laid out as each new candidate enters the assessment room, to be standardised as much as possible. It also provides a recording of the candidate performance which is helpful should there be any issue raised about the conduct of the exam task or to verify if an unexpected event eg a light bulb blowing, happened during the assessment. Of course, anything unusual would normally also be captured by the assessor in their incident log.

The provision of tailored support for ICER candidates was, rightly, of the highest quality, with appropriate adjustments made to enable these candidates to have the best opportunity to demonstrate their ability on a par with candidates without the need for exceptional requirements.

The candidate briefings were very clear, helping candidates to fully understand what to expect on the day and what they were expected to do or not allowed to do. In the sessions I observed there were few questions asked by candidates, which I took to be a testament to the clarity of the presentation.

Overall, the conduct of the day was deeply impressive with excellent facilities at the RCGP building in Euston being used**. My only recommendation is to maintain the high standards of delivery demonstrated on this day in the live SQE2 context**.

5 Pilot performance data

5.1 The functioning of SQE2 pilot tasks/assessment stations

In the SQE2 pilot, candidates took 14 tasks or stations in total. Kaplan provided the performance data for each station, the seven core stations and the seven optional business or criminal stations. A wide variety of mean scores were achieved by candidates, the highest being 82.7 out of 100 for the business interview station, the lowest was 41.75 for the criminal writing station. Standard deviations for each station ranged from 26.85 on the business case and matter analysis station to 14.61 on the core attendance note station. In general, standard deviation scores were higher on the application of law scores than the skills scores and the mean was generally lower for application of law scores. This suggests a wider range of performance in appropriate application of law and SQE2 pilot candidates generally found this more challenging than the skills-based elements of the tasks. Overall, Kaplan reported that pilot candidates displayed a wider range of performance than would be expected in a live context, with a longer tail of poor performance and some extremely good responses. This was suggested as being due to two factors. Firstly extremes of motivation for pilot candidates, some being very motivated to do well because this would bolster their career prospects if they are able to report pilot outcomes to their firms and others for whom little rode on the outcome and therefore they did little or no preparation. Secondly pilot candidates had not already passed SQE1 which will be a pre-requisite for sitting SQE2. This atypical performance in the pilot and having 167 candidate responses must be considered when the pilot results are analysed. It means final SQE design decisions should not be based solely on the SQE2 pilot outcomes. They are helpful and indicative but should not be over interpreted or solely relied upon for policy making.

Kaplan produced draft psychometric analyses about the performance data for the SQE2 pilot in March 2020. These provided a thorough and detailed overview of: candidate demographics; the performance of the multiple choice tests (MCT) and the assessment stations.

The MCT were used to enrich the analysis of the legal skills assessment stations data and are not intended to be used in the live SQE2. All candidates took 60 questions sampling across all the areas covered in SQE2 (core) and then a further 60 questions in their chosen specialism – business or criminal. The candidate performance on these tests demonstrated the very wide range of candidate ability and unusual performance of higher and lower scoring candidates, which provides evidence of unusual extremes of response. The candidates who selected the business specialism outperformed the criminal candidates on core items, whereas the criminal candidate outperformed the business candidates on their specialist items. These data offered a useful reference point for candidate performance on the skills stations.

Although detailed candidate demographic data were provided, given the small numbers in the sub-groups of candidates once most of the data categories are separated, it is unwise to draw any strong conclusions from these data.

Very detailed item performance data across the skills assessment stations were provided by Kaplan. Overall, these demonstrated the skills stations performed well, discriminating effectively. Candidates performed better on the skills elements compared to the application of law elements of the marking criteria. While all stations functioned effectively the standard deviations for the interview (which is the only task to solely assess skills and not application of law) and attendance note stations were generally lower. Kaplan carried out double marking of at least one station from each Legal Practice area and each skill. In general, these data demonstrated good standards of marking and suggest standardisation of markers was effective.

A range of univariate and multivariate analyses were provided by Kaplan. The univariate analysis showed differences in performances by individual candidate variables eg home background, work experience, legal experience and personal demographics. The multivariate analyses aim to explore which are the best true predictors of candidates' total score. These analyses need to be viewed with caution given the small numbers which make up these data sub-sets and complexity of the analysis. The outcomes from these analyses were unsurprising, for example, as would be expected, on the multivariate analyses higher MCT scores was the best predictor of overall performance. The analyses were not conclusive about the value of prior work experience being a strong predictor of overall performance, there was a slight indication it was of some benefit.

5.2 Key issues arising from the SQE2 pilot performance data

The issue of the extent to which candidates can compensate for poor performance on one or more skill by achieving higher scores on others and still pass overall was considered in the Kaplan analyses. Overall candidates displayed fairly uniform performance across skills stations, with limited exceptions on the attendance note and interview stations. A particular focus of the analyses was the advocacy stations, advocacy has special status because it is a protected characteristic of solicitor qualification. The evidence showed a very low risk that candidates would be able to pass overall while performing poorly on the advocacy, or any other, station. So while some limited compensation between skills did occur it was not pronounced. **I recommend when finalising the SQE2 exam this evidence is considered when exploring ideas for the composition of live SQE2 stations to reduce this low risk even further**. This is why Kaplan's suggestion to combine the interview and attendance note in to one station in the live SQE2 is worth exploring, as is their recommendation to have more assessment stations overall.

The issue of the weighting of the skills and application of law as part of the skills stations marking criteria was considered. In the SQE2 pilot this followed a 50/50% split. Kaplan have recommended maintaining this split for the live SQE design, based on evidence from the pilot that moving this ratio to 60:40 or 70:30 in favour of skills would increase (60:40 – slightly, 70:30 – not so slightly) the likelihood a small number of candidates pass when achieving low scores on application of the law, an issue that had been previously been considered in the QLTS context where application of law was increased to a 50:50 ratio. **Considering this pilot evidence, I recommend that the final design should be a ratio 50:50 or 60:40 skills versus application of law.**

Prior to the pilot taking place Kaplan had predicted that 14 stations were likely to be at the lower end of the number of stations that would produce an acceptable standard error of measurement (SEm). This is supported by the indicative pilot outcomes. **I support Kaplan's recommendation to have more than 14 stations in the final SQE2 design**, although not substantially more, it will provide slightly greater reassurance around reliability of overall results by slightly reducing SEm to a lower and more defensible level. Too many more will have diminishing returns around reliability and make SQE2 more complex, expensive and potentially overly demanding for the candidate, so a balance needs to be struck.

The most significant issue to resolve for the final SQE2 design is whether it should be a uniform model (where all candidates take the same skills stations), which take a sample across all areas, or a core plus specialism model (as was the case in this pilot) or allow candidates to choose which two specialisms to be examined in. The overwhelming evidence from a psychometric perspective is to have a uniform model. This is backed up in the relevant academic literature and is the model usually followed by professional qualifications which lead to licensure. From a defensibility perspective, and in order to ensure fairness to all candidates, the evidence from the pilot is the SQE2 design should be uniform, and while recognising this will not be some key stakeholder's preference, the design must be able to withstand any reasonable challenge. The evidence provided by the pilot demonstrated the difficulties with equating or creating a common scale, using a common core as a proxy for the candidate cohorts across just two specialisms. It showed the two specialisms had candidate cohorts which were not similar enough, the common core was not equally representative of both specialisms. This meant it was not a mini-version of the tests being equated and therefore could not effectively act as a yardstick for performance. It could not safely anchor standard setting across the specialisms. If this is the case for the SQE2 pilot, where there were just two specialisms, this issue is likely to be much worse where five specialisms would be available in a live context and worse still if candidates could pick any two specialisms from the five available. This problem would be exacerbated by the predicted low numbers of candidates in the early years on the SQE qualification because each sub-set by specialism choice would not have sufficient numbers to make their data statistically significant or reliable for standards setting.

This core plus specialism model would only be possible if there were five separate qualifications or endorsements, allowing each of the five specialism choices to have its own pass mark for each cohort. This of course goes against the spirit of an overarching solicitor qualifying exam. It also makes the option of complete candidate choice indefensible unless each pair of specialisms was also treated separately, in effect 10 different qualifications, which is even more unpalatable.

I recommend the final SQE2 design is a uniform model. In making this recommendation this means candidates that have received more training through classroom teaching will not be able to be narrowly drilled in a set of skills in one or two specialisms. The unpredictable nature of a uniform exam, which samples across all areas, prevents this gaming type strategy, protecting the integrity of the exam. I recognise there are concerns the uniform model is likely to lead to training providers and candidates concluding they will need more intensive preparation for SQE2 which creates concerns about the training time and costs for candidates. This is because they may decide that a uniform model requires them to cover, or at least refresh, the knowledge or application of law, as covered in SQE1. This is why I have recommended that SRA and Kaplan review how to reduce the preparatory burden on candidates.

I recognise there are still concerns about whether work experience across all five areas would be needed in order to best prepare for a uniform SQE2 exam. It appears that work experience in one specialism is likely to offer some assistance to candidate preparation, but so far there is no evidence that work experience across all five areas would be necessary. Indeed it should not be a requirement for candidate preparation, this would be unrealistic and too disruptive for firms and candidates alike. Work experience and/or some types of classroom training is likely to help candidates prepare for SQE2. **In the first few sittings of the new SQE, I recommend the effect of candidate preparation, with regards to the value and nature of prior work experience, is monitored.** Of course this will be self-declared by the candidate so will not be entirely objective.

Overall the quality of the Kaplan SQE2 pilot data analysis was very high, and this excellent level of analytical and psychometric support should become an ongoing hallmark of the roll out of the SQE.

6 Summary of recommendations for SQE2

6.1 The training and standardisation of actor assessors showed much best practice and these high quality activities should be sustained in the live SQE2.

6.2 When selecting performance exemplars for standardisation for the live SQE2 a wide spread of overall performance on the global marking scale is needed, 'marginal fail' and 'marginal pass' exemplars are always important to include.

6.3 Kaplan should video record the standardisation role play activities and make these videos available to assessors after the standardisation and training days to assist consistency of application of the marking criteria.

6.4 Kaplan have shared their high-level plans to offer training and support on how to avoid unconscious bias in the future, amongst a comprehensive range of other measures, to reduce the risk of any unfairness to any candidate as a result of their ethnicity, religion or background. It is important these plans are executed.

6.5 When allocating scores, it is important assessors understand the importance of, and difference between, applying a domain and global score for each assessment. Kaplan should ensure there is always an explanation to assessors, especially any

new assessors, about how the pass mark is set and the impact of their judgements in the live context on standard setting.

6.6 Narrative advice for assessors, to sit alongside the published marking criteria descriptions, about how to form judgements in the live SQE2 was provided to assessors in the SQE2 pilot. For the live SQE2 this narrative advice should fully clarify how to interpret the 'application of the law' marking criteria when making domain scoring judgements, which is especially important for any new assessors.

6.7 The provision of materials to support assessment tasks needs very careful consideration to strike the right balance between ensuring this is a skills-based assessment, without candidates needing to remember lots of detailed legal knowledge, while expecting basic legal concepts and application of the law to be understood and applied. The advice made available to candidates preparing for the SQE2 must be clear about what are the legal concepts and knowledge they are expected to know. Most importantly all options for reducing the preparatory burden for candidates need consideration. It is entirely appropriate that some tasks do, and some do not, have supporting materials, according to the objectives of that assessment task and its construct.

6.8 The conduct of the candidate assessment day was deeply impressive with excellent facilities at the RCGP building in Euston being used. The high standards of delivery demonstrated on this day should be maintained in the live SQE2 context.

6.9 When finalising the type and number of tasks for the SQE2 Kaplan's suggestion to combine the interview and attendance note in to one station in the live SQE2 is worth exploring.

6.10 The final design should have a ratio of 50:50 or 60:40 skills versus application of law for each assessment station that assesses both.

6.11 There should be more than 14 stations in the final SQE2 design, but not too many more (I would suggest a maximum of 18).

6.12 The final SQE2 design should be uniform.

6.13 In the first few sittings of the new SQE, the effect of candidate preparation, with regards to the value and nature of prior work experience, should be monitored.