SQE Stage 1 pilot findings June 2019

Geoff Coombe: SQE Independent Reviewer

Crickcombe Ltd

1. Executive Summary

Overall the SQE Stage 1 pilot, held in March 2019, successfully achieved its purpose. This was to explore whether SQE stage 1 is a fair, reliable, accurate, valid and cost effective and manageable assessment. The questions raised in the assessment specification about the SQE stage 1 pilot and my recommendations are in section 4.

The pilot demonstrated what was likely to work well, or not so well, in the live context and allowed the planned assessment methods to be effectively tested by over 300 well-chosen pilot candidates. This has provided invaluable data and experience which will inform the design of the final SQE Stage 1. The Functioning Legal Knowledge (FLK) assessments performed well, however the performance of the pilot Stage 1 skills assessment raised significant concerns about reliability, fairness and standard setting.

2. Functioning Legal Knowledge (FLK)

2.1 The functioning of the pilot questions: reliability and validity

The questions used in the Stage 1 pilot comprised of 360 objective test items (individual questions), which required a single best answer, of which 176 were new items designed specifically for the pilot and the remainder were written for QLTS tests, many of which were re-used. Item writing for the pilot was conducted using experienced writers who are all solicitors of England and Wales and collectively have experience that ranges across teaching as well as practice. Items were then edited, through multiple points of review, including by qualified solicitors of England and Wales with experience of editing objective test questions for exams leading to licensure. A sample (approximately 10%) of the edited and reviewed questions were sent for review to a US qualified lawyer with experience of editing questions for the US Multi-State Bar Exam. Kaplan have an experienced team of in-house reviewers, who are also solicitors and are skilled in compiling and formatting objective test items ready to be presented to candidates on-screen. The external editor who reviewed pilot questions has also worked with Kaplan on QLTS for several years.

Kaplan have committed to future items also being sent to practitioners, not involved in writing or editing, for them to comment on their relevance to practice, which will provide an added safeguard to ensure validity. Overwhelmingly, test items performed well and re-using items from QLTS which are used in a very similar, but not identical, context had the advantage of knowing prior candidate performance on these items.

Having met Kaplan to discuss the item writing process and their reflections on how to improve it further for the live context I was satisfied that the methods used should lead to good items being created for live tests. My recommendations are:

2.1.1 Kaplan were able to explain in detail the processes they follow to recruit, train and standardise those writing items for the FLK, as well as blueprinting, editing, review and test rendering for display to candidates. **These processes should be fully**

documented ready for the live context. The documentation should adhere to a recognised quality management standard so it can be used for audit and continuous improvement purposes.

2.1.2 There is a reliance of a small number of key highly skilled individuals within Kaplan to oversee the test creation and rendering process, therefore a risk management plan with **effective countermeasures to deal with the potential non-availability of key test creation staff** is needed.

2.1.3 As part of the review of future live item performance Kaplan **should use an independent expert to review test analyses with their academic team**. This will provide additional objectivity and assurance when making decisions such as whether to include an item that has performed unusually in a live test.

Across the three 120 question FLK tests, candidate marks ranged from 17% to 85% of the total marks available. Mean scores and standard deviations (which shows how much group candidate performance varied from the mean score) on each test were very healthy and other reliability statistics were generally good. Kaplan conduct a thorough review process where various item data analysis methods are used to highlight any question that has performed unexpectedly. As would be expected in any test of this sort there were a small number of items where improvements could be made if similar question types were used in the future. There is careful analysis of the performance of each item, using a variety of psychometric techniques, to ensure lessons are learned from each examination cycle. These analyses included: the popularity of the incorrect answer; candidate response time to each question; item facility; item total correlations and reviews into unusual candidate response patterns.

If an answer produced, for example, more candidates selecting an incorrect than a correct answer, such questions received special attention. Firstly to check that the question as posed was valid in the view of experts, and therefore legitimate to include in the test, then to assess whether aspects such as level of difficulty, or structure of the question, should be amended if similar questions were used again for future tests. During this review no pilot test items needed to be removed from the test due to them being flawed, however a few items will be retired from future tests or amended. In a number of cases, trends or themes of poor candidate performance were demonstrated. It would therefore be useful to share these themes with future candidates to help support their preparation for live tests.

Pilot candidates were asked to provide feedback on their experience of the pilot tests. This feedback provided no significant issues which would raise concerns about the validity of any items, neither did the feedback from solicitors used by Kaplan at the Angoff panel.

2.2 Angoff and standard setting

An Angoff method for recommending a pass mark on the FLK was used. Angoff is a method that uses a panel of experts to judge how difficult each item is, to help to determine the cut score (pass mark). I observed the Angoff meeting, and overall the decisions made, and outcomes derived, were sound and certainly satisfactory for a pilot context. The Angoff panel comprised solicitors of England and Wales who were all familiar with the standard of a day one solicitor, they came from a range of backgrounds and experiences, including teaching and practice and some newly qualified solicitors. In addition to some calibration items the panel made judgements about the level of difficulty of all 176 new items created for the pilot. Seven of the nine panel members were able to stay to review 24 QLTS questions, used in the most recent sitting of the QLTS, which were used for comparison purposes. Prior Angoff judgements about other re-used QLTS items were carried forward as standards for common items were statistically similar.

The experts worked hard and diligently through a long day, however some struggled to complete the task in the time allotted and stayed on beyond the planned finish time to complete the task. The first Angoff meeting for the first set of live FLK tests will be particularly important because it will set a standard which would normally be carried forward for subsequent FLK tests. In the light of pilot experience some improvements can be made to this process when it is conducted in the first live setting. I have the following recommendations:

2.2.1 For the first live SQE1 Angoff panel all 360 SQE1 FLK test items should be judged, even if any are being reused from a QLTS setting. The context between QLTS and SQE, while very similar, has differences and all items should be judged against the SQE standards expected for a Day 1 newly qualified solicitor.

2.2.2 The solicitors invited to the first live SQE Angoff panel will need to carefully selected to have not just relevant legal knowledge, but also knowledge and experience of what can be expected of a Day 1 qualified solicitor, as well as **having the significant time available to dedicate to this task.**

2.2.3 Kaplan will need to **schedule a longer duration to conduct the first live Angoff** meeting.

Overall the Angoff process was suitable for this pilot setting and the outcomes derived should be secure.

I observed the pre-award meeting, which rehearsed how recommendations for pass marks for the 3 x 120 question pilot FLK tests could be made. While pass mark recommendations were made at the pre-award meeting, a cut score (pass mark) was not set at the subsequent final award board. This is because it was not necessary for the purposes of the pilot. However going through this process enabled invaluable learning for the live context.

The pre-award meeting was thorough in its consideration of the relevant issues and carefully explored a range of approaches which could be taken. The data analyses provided for this meeting were very good and a well-informed discussion about the implications of where to set the pass marks took place. For example, the discussion about standard error of measurement (which is a feature of all assessments in any context) and how to take account of this when deciding the pass mark, was excellent. The recommendation from the meeting leads to a very low risk that any candidate achieves a pass on the FLK without deserving to do so. This will influence pass rates in the live context if the same approach is carried forward. Essentially the decision on where to place the cut score balances competing risks around too few 'right' candidates, those that narrowly deserve to pass, and the 'wrong' candidates, those that narrowly don't deserve to pass. In the first live setting, the SRA exam board need to ensure they are comfortable with this trade off, the recommendations from the pre-award meeting, if carried forward for the first live context, set a demanding standard for candidates.

Ultimately standard setting should meet these criteria: be fair, defensible and command public confidence. I recommend:

2.2.4 SRA should review the approach to the standard setting process and parameters used, such as treatment of standard error of measurement, both after the SQE 2 pilot and in the first live setting, to ensure these standard setting criteria are met. This review should also ensure that the live overall SQE outcomes (when all components are aggregated) support the definition of the minimally competent candidate as defined in the assessment specification.

2.2.5 The analyses provided and the discussion at the pre-award meeting were excellent. During this discussion it was agreed that the SRA and Kaplan will establish formal **data protocols and schedules, awarding policies and processes and agree what information to send to participants prior to, and provide at, the pre-award and award meetings.** These should be documented to aid transparency, assist audit and continuous improvement and support defensibility.

2.3 Pilot candidate and stakeholder feedback

Kaplan conducted a feedback survey of all candidates that took part in the pilot. In addition, Kaplan and SRA have held feedback events and meetings with key stakeholders to discuss views on the Stage 1 pilot. A few social media and on-line and press based media articles about the pilot have been published and these have offered a range of feedback both positive and negative. One view that did emerge, but was by no means universal, was that the FLK questions were too easy. The evidence from the performance of the candidates in the pilot suggests this was not the case, with around only 50% of pilot candidates achieving over half marks. While the performance of pilot candidates will not reflect the performance of live candidates, who should be better prepared and more motivated, it is very unlikely the first live cohort of candidates would find these tests too easy if broadly equivalent levels of item difficulty are maintained.

2.4 Number of FLK tests for Stage 1 in the live context

Kaplan have recommended 2 x 180 item FLK tests for the live SQE, rather than 3 x 120 item tests. This is because they believe this will improve the reliability of these tests and therefore make the outcomes for SQE more robust, for example to potential challenge about incorrect outcomes. Some stakeholders have expressed concern about the potential for candidates to adopt revision strategies which might allow them to pass the FLK without appropriate depth and breadth of functioning legal knowledge in some important areas. Therefore, concerns have focused on the opportunity for candidates to be able to compensate for poor performance on some key areas of FLK by thoroughly revising, learning and understanding others and still achieve a pass. Kaplan produced some excellent data analyses to review this concern. The evidence from the pilot demonstrates that candidates would find it very difficult to achieve a pass without good functioning legal knowledge across all areas, partly due to the integrated nature of some of the questions which require knowledge from more than one area of FLK. It is possible, when creating the blueprint for future tests, to more heavily test areas of FLK if they are considered more important than others. But there would need to be a validity based argument for doing so. Changing the weighting of areas of the test would mean more questions are aimed at these topic areas, which would further reduce the already low risk that strategic revision (which purposefully neglected a topic of the FLK) would be successful in achieving a pass.

The pilot SQE has several components, or hurdles, which need to be passed to achieve an SQE pass. These are deliberately set to ensure that outcomes deliver a high standard and each component is robust and rigorous, demonstrating the standard required of a Day 1 solicitor. However the more hurdles an examination contains, for which each component must be passed, the fewer candidates that will achieve an overall pass. Therefore the design of the new SQE needs to be carefully balanced to be rigorous, set the right standards for each component, while delivering outcomes which command the confidence of candidates, learning providers, firms and end users of the qualifications and the wider public. There is a risk that having three rather than two FLK tests in Stage 1 will suppress pass rates to an unacceptably low level. While stakeholder concerns about candidates being able to compensate good FLK in some areas against poor FLK in others are understood, the evidence from the pilot suggests this risk is low. And with an appropriate blueprinting strategy this risk can be further mitigated. Kaplan understand the need for, have experience of, and are planning, a thorough and detailed blueprinting strategy for the live FLK to support a high-level, public facing document.

I therefore recommend:

2.4.1 There should be 2 x 180 item FLK tests in Stage 1.

2.4.2 SRA review the weighting and relative importance of the topics in the final FLK syllabus content and work with Kaplan to publish final wording of the syllabus content. Small improvements to the wording also helps to address some suggestions to improve clarity that Kaplan question writers have noted.

2.4.3 Kaplan continue to create a detailed blueprint for each live FLK test, for internal use, supplementing the high level published blueprint, which enables: appropriate weighting of each topic, major headings and sub-categories of FLK to be profiled; all aspects of the syllabus to be tested over time and without it being possible for candidates or learning providers to successfully predict future test questions.

2.4.4 Candidates and learning providers are informed of the expected weighting of FLK topics in the live tests, through the high level blueprint.

2.4.5 A narrative summary of feedback on pilot candidate performance is published and shared, highlighting patterns of poor and good performance observed when answering pilot FLK items, to enable better preparation by live candidates.

3. Stage 1 Skills

3.1 The functioning of the skills assessments: reliability and validity.

The skills part of the Stage 1 pilot comprised one legal research task and two writing tasks. For the pilot, and to aid decision making, candidates were required to complete this cycle twice. Unlike the FLK, these tasks are assessed at a lower level, a level aimed at testing a candidate's readiness to commence legal practice in an unqualified capacity. There were two Objective Structured Clinical Examinations (OSCEs), each containing one research and two writing exercises in two different contexts. OSCE is a well-established and proven method of assessment, usually used in clinical related healthcare settings where high stakes practice and observation skills need to be assessed by expert assessors.

There were three stations for each ‘mini’ OSCE and the marking criteria was a six-point scale (A - F), A worth 5 marks, F worth 0 marks. The two assessors used also provided a global score, to assist standard setting, of either fail, marginal fail, marginal pass or pass. Marks achieved in the OSCEs ranged from 8 - 100% of total mark available. Each station for each candidate was marked by the two assessors. Overall reliability scores were reasonable for an assessment with such a small number of stations. These were however, lower than would be expected or considered satisfactory in a live high-stakes OSCE setting. These low levels of reliability are mainly a function of the limited number of stations and narrow range of marks available for this type of skills assessment. In the pilot context it was possible for two assessors to produce relatively similar marking outcomes for the same station and task, although even with just two assessors there were still noticeable differences in mean scores for two stations on the mini-OSCE 2. The agreement between markers on the global scores was satisfactory, albeit working to a standard which was difficult to define.

Taking this evidence into consideration there is a risk that when more markers need to be used in a live context, this no better than satisfactory level of consistency will be difficult to improve and therefore consistency of marking will be lower. Overall standard errors of measurement were higher than what would be considered acceptable in a live setting. It was also noted that candidates achieved much lower marks on one of the stations. This was reviewed by the markers involved and Kaplan who concluded these lower scores could be explained by candidates generally finding tax related tasks difficult, not that the task was flawed. If this component was to be continued in a live setting, an independent view on validity would offer further assurance.

Detailed analysis suggests the nature of the skills being assessed, twinned with the style of assessment, appears to have advantaged certain candidate groups, even when taking the same candidate performance on the pilot FLK tests into consideration. The outcome of this pilot means that inclusion of these types of tasks in the Stage 1 SQE threatens the likelihood of progress of certain candidate groups. Worst still it could be argued that inclusion of this nature and style of assessment in the live SQE raises the risk of discriminating against some candidate groups.

From a validity perspective, having an assessment as part of Stage 1 which is at a different, lower, but therefore difficult to define, level to the overall SQE standard (exemplifying that needed of a Day 1 newly qualified solicitor) is problematic. This confuses the definition of the overall SQE minimally competent candidate. This is because this component of the exam confuses the overall aim and overall standard of the exam and makes standard setting across the whole SQE inconsistent.

3.1.1 For these reasons of reliability and validity I recommend that the **Stage 1 skills component is removed from the live exam**. There should be other methods outside the exam for which candidates and firms can assess these skills as part of a trainee's development, and SQE2 does require some of these skills to be assessed at the full SQE standard.

3.1.2 Kaplan should review the reason(s) for differences in marking across assessors on two stations on mini-OSCE 2, to review if any **lessons need to be learned to assist preparation for the SQE2 pilot with regard to advice and training given to support assessor standardisation when marking**.

3.2 Standard setting

Despite the reservations outlined above, because the stations were doubled marked and reliability scores overall were, just, satisfactory in the circumstances, it was helpful to use the pilot candidate scores in a pre-award setting process in order to pilot these. Once again the final cut score was not set at the subsequent final award meeting because it was not necessary for the purposes of the pilot.

Kaplan's report provided for the pre-award meeting details the method used. At the pre-award meeting a good discussion about the various methods that could be used to determine the pass mark took place. The conclusion of which was to use borderline regression, which included the global assessments and scores for each station and combined with a recommended standard error of measurement level consistent with that recommended for pilot FLK (+1.64 SEm) to provide a standard for the whole skills assessment. I support this recommendation from a consistency argument, but is should be reviewed during the SQE2 pilot.

4. Summary of questions asked of the pilot from the assessment specification and recommendations

4.1 What is the minimum number of assessments required in SQE1 to reliably and validly assess the FLK? What number, format and type of questions and length of assessments will most reliably and validly assess the FLK?

I recommend 2 x 180 mark FLK tests only, and I recommend not including the skills assessment in the live SQE1. The evidence from the SQE1 pilot demonstrates objective tests, requiring a single best answer, are well suited to reliably assess functioning legal knowledge and I have no concerns about validity based on evidence and all forms of feedback from the pilot.

4.2 What should the balance be between SQE1 and 2? For example, can legal drafting be reliably and validly assessed at SQE1?

We are yet to pilot SQE2, so this question should be reviewed after the SQE2 pilot later this year. However, the standard of, type, and number of skills assessments used in this SQE1 pilot leads to a recommendation not to include skills assessments in SQE1 for the reasons mentioned in section 3.

4.3 What are the benefits and risks of a non-compensatory standard setting model as opposed to total compensatory or partial compensation?

It is usual for high stakes, professional licensure-based examinations to be made up of non-compensatory components. When designing such an examination for the first time it is very important to investigate and model what each component or ‘hurdle’ contributes ie how essential it is, and how, by including it, it will impact overall qualification pass rates. I recommend removing the skills assessments from SQE stage 1, partly because it confuses the overall SQE standard and also because it is unlikely to deliver a sufficient level of reliability or validity in a live context. I also recommend 2 x 180 FLK tests rather than 3 x 120 mark FLK tests, partly because there is evidence from Kaplan’s analyses this could improve reliability but mainly because having one fewer hurdle is likely to mean more accurate pass/fail decisions to ensure that those who pass deserve to pass and fewer candidates that (narrowly) deserve to pass do not fail. It is worth remembering that the risks of candidates successfully passing these tests, whether at 120 or 180 marks, without good knowledge across all FLK topics, seems small according to the analysis conducted by Kaplan and with an effective blueprinting strategy this risk further declines.