The Solicitors Qualifying Examination Pilot of the Assessment of Functioning Legal Knowledge: a psychometric and statistical analysis

Authors

Case SM[1], Coombes LR[2], Fry E[2], Swanson DB[3], Wakeford R[4]

Affiliations

1. Formerly, National Conference of Bar Examiners, Madison, WI
2. Kaplan, Spring House, 40-44 Holloway Road, London
3. American Board of Medical Specialties, Chicago, IL
4. Hughes Hall, University of Cambridge, Cambridge

* * *

EXECUTIVE SUMMARY

A key component for a proposed route to qualification as a solicitor of England and Wales is an assessment of the application of candidates' functioning legal knowledge (FLK) within the planned Solicitors Qualifying Examination (SQE). Kaplan was appointed as the assessment provider by the profession's regulatory body, the Solicitors Regulation Authority (SRA), to design, develop and deliver the assessments. Since 2011, Kaplan has provided a qualifying exam of both applied knowledge and legal skills assessments for lawyers qualified in other jurisdictions and barristers of England and Wales to become solicitors of England and Wales (the qualified lawyers transfer scheme). This paper provides details of a pilot delivery of the proposed SQE assessment of FLK.

Three papers, each of 120 single best answer, multiple choice questions (MCQ) were constructed as follows: 1) Business Law and Practice, Dispute Resolution, Contract and Tort; 2) Property Practice, Land Law, Wills and the Administration of Estates and Trusts, Solicitors' Accounts; and 3) Public and Administrative Law, the Legal System of England and Wales; Criminal Law and Practice. The potential for two papers of 180 items each was also evaluated. A total of 316 volunteers took the three tests and provided demographic information.

The quality of test items was assessed using indices of difficulty, response time and corrected item-total correlation (discrimination) supplemented by analyses based on item response theory (IRT). Candidates' test scores were calculated for each of the three papers and for the three papers as a whole (360 items). Overall scores were examined against individual demographics and then using multivariate analysis to identify the key predictors of performance. Subject scores were correlated against paper scores to assess the extent of compensation between subjects. Data on the reproducibility of test scores (reliability using coefficient alpha, and precision using the standard error of measurement (SEm)) were calculated for the three 120-item tests, and, using generalisability theory, modelled for different test lengths and evaluated against current best assessment practice.

Reliable and accurate assessment of the application of the FLK using single best answer items was achieved in the Stage 1 pilot. In general, the items showed appropriate statistical characteristics. Item review by solicitors was found valuable to complement traditional interpretations of item statistics. Content expert review following the exam administration found that there were gaps in knowledge (reflected in some items that candidates found very difficult). There was only modest room for item improvement overall. Three 120-item tests did not quite achieve the robustness of scores necessary in such a credentialing examination. However two exams of 180 items would reach

these levels. Candidate feedback was encouraging. Initial tentative indications are that educational factors (having successfully completed the Graduate Diploma in Law (GDL) and having undertaken a law degree in a Russell Group university in England and Wales) were key predicators of success in the exam. While this provides concurrent validity for the exam, it is also a cause for concern, given the confounding of these educational factors with membership of minority groups protected under the Equality Act 2010. Industry-standard targets of reliability and precision can be achieved with two tests of 180 items, each with a separate pass/fail point and with only limited compensation occurring between subjects within a paper.

* * *

INTRODUCTION AND BACKGROUND

Background

In 2021, the Solicitors Regulatory Authority (SRA) will introduce the Solicitors Qualifying Examination (SQE) as a single point entry for those wishing to be licensed as a solicitor of England and Wales[1]. Following extensive consultation with stakeholders, the SRA Board decided in April 2017 to introduce the SQE. This was followed by publication of a draft Assessment Specification in June 2017[2]. A year-long tender to design, develop and deliver the SQE was won by Kaplan in August 2018[3]. Kaplan has, since 2011, provided a qualifying exam of applied knowledge and legal skills assessments for lawyers qualified in other jurisdictions, and barristers of England and Wales, to become solicitors of England and Wales – the qualified lawyers transfer scheme (QLTS)[4].

The SRA envisage that the SQE exam will have two parts. SQE1 will principally focus on application of core legal knowledge – the test of Functioning Legal Knowledge (FLK). It is undecided if SQE1 will also include a legal skills test and this issue is not addressed further in this paper. SQE2 will involve the assessment of candidates' practical legal skills. The FLK test—the focus of this paper—would utilise single best answer, multiple choice question (MCQ) items, shown to be sufficiently accurate and reliable for assessing applied legal knowledge in the QLTS assessments[5,6].

A Pilot Delivery of the SQE1 Functioning Legal Knowledge Assessment

As part of the development process, and following further stakeholder engagement and expert analysis, an Assessment Specification for an SQE1 pilot was published in 2019[7]. To provide acceptable sampling across the FLK, a total of 360 items were projected as sufficient. These would be split into three separate papers blueprinted to the FLK[8] .

[1] https://www.sra.org.uk/sra/policy/sqe/
[2] https://www.sra.org.uk/globalassets/documents/sra/news/sqe-draft-assessment-specification.pdf
[3] https://kaplan.co.uk/insights/article-detail/insights/2018/08/01/kaplan-appointed-as-sqe-assessment-organisation
[4] https://qlts.kaplan.co.uk/home
[5] Eileen Fry, Jenny Crewe & Richard Wakeford (2013) Using multiple choice questions to examine the content of the qualifying law degree accurately and reliably: the experience of the Qualified Lawyers Transfer Scheme, The Law Teacher, 47:2, 234-242
[6] Eileen Fry & Richard Wakeford (2017) Can we really have confidence in a centralised Solicitors Qualifying Exam? The example of the Qualified Lawyers Transfer Scheme, The Law Teacher, 51:1, 98-103
[7] https://www.sra.org.uk/sra/policy/sqe/pilot/sqe-assessment-specification
[8] https://www.sra.org.uk/sra/policy/sqe/pilot/sqe-assessment-specification Annex 3

Each of the papers was aligned to areas of the blueprint in the Assessment Specification. The first of these (Paper 1) assessed application of Business Law and Practice, Dispute Resolution, Contract, and Tort; Paper 2 covered Property Practice, Land Law, Wills and the Administration of Estates and Trusts, Solicitors Accounts; and Paper 3 covered Public and Administrative Law, the Legal System of England and Wales, Criminal Law and Practice. Ethics was assessed pervasively across all subject areas. Table 1 shows the blueprint that was followed.

Table 1 : FLK assessment Blueprint

Paper 1 **Business, Dispute Resolution, Contract, Tort**	**Assessed as a % of paper**	**Paper 2** **Property, Wills, Solicitors Accounts**	**Assessed as a % of paper**	**Paper 3** **Public law, Legal System, Regulation, Criminal**	**Assessed as a % of paper**
Business organisations, rules and procedures, including taxation of business organisations	20-30%	Core knowledge areas of freehold and leasehold real estate law and practice, including core principles of planning law and property taxation. Solicitors Accounts in the context of conveyancing.	20-30%	The Legal System of England and Wales, Sources of Law, Constitutional and EU law, the Human Rights Act 1998 and the Equality Act 2010	25-40%
The principles, procedures and processes involved in dispute resolution	20-30%	Core principles of land law	20-30%	Regulation: Money Laundering and Financial Services. Legal Services	15-20%
Core principles of contract law	20-30%	Wills and Intestacy, and Probate and Administration Practice. Taxation and Solicitors Accounts in the context of Wills and Probate Practice.	20-30%	The procedures and processes involved in advising clients at the police station and in criminal litigation	20-30%
Core principles of tort	20-30%	Core principles of trust law	20-30%	Core principles of Criminal Liability	20-30%
Ethics	Pervades	Ethics	Pervades	Ethics	Pervades

The method of classification used for this blueprint is 'single best classification', meaning that an item is assigned to the classification category that is considered to be the primary focus of that question. This does not imply that questions are discrete and neatly fit into each category as, for example, an item may require knowledge of both dispute resolution and contract in order for a candidate to answer it correctly. As such, the blueprint describes an assessment as having 20 percent of questions on a particular classification, but application of the knowledge in that classification would be required to answer more than 20 percent of the questions included in the paper.

The pilot provided an opportunity to test the creation, delivery and analysis of the assessment modality. It also particularly aimed to assess how many test items are required to meet the minimum psychometric standards required for a high-stakes licensing assessment in law; also the extent to which an examination covering several different areas of law assured adequate knowledge of the component subjects.

METHOD

The Candidates for the Pilot

The candidate recruitment process was conducted via an application form on the SRA website and resulted in 612 applications to take the pilot assessments. Applications were encouraged from minority groups protected under the Equality Act 2010.

Candidates were selected to be, as far as possible, representative of those who are thought likely to sit the SQE. In terms of prior education and experience this meant candidates were selected who:

- had completed Stage 1 (the compulsory stage) of the Legal Practice Course (LPC); or
- had completed a period of study and/or work experience equivalent to Stage 1 (the compulsory stage) of the LPC; or

- were qualified lawyers in a recognised jurisdiction eligible to qualify via the QLTS; or
- were barristers of England and Wales.

555 candidates were invited to take part in this SQE1 pilot. 419 of the invited candidates accepted their place. 58 of them cancelled in the run up to the examinations and there were 43 no-shows on the day of test administration, leaving 318 active participants. Two of these did not sit the full FLK assessment and are excluded from this analysis. This report, therefore, presents demographic data and analysis of performance on the FLK for the 316 candidates who attended all three FLK assessments.

A summary of the 316 candidates' demographics is shown in Table 2 (a, b & c), and covers their home background and work experience (2a), legal education (2b), and personal demographics (2c).

Table 2a SQE1 FLK Assessment: Pilot Candidate Demographics

Home background and work experience

Variable	Value	N	%
Level of parental education when candidate was 18	1+ degree level qualification	143	45.3%
	No formal qualifications	67	21.2%
	Not stated	7	2.2%
	Qualifications below degree level	99	31.3%
	Total	316	100.0%
School type classified	Independent	33	10.4%
	Other / not stated	7	2.2%
	Outside UK	70	22.2%
	State: non-selective	151	47.8%
	State: selective	55	17.4%
	Total	316	100.0%
Experience as CILEX, trainee or apprentice solicitor, or paralegal?	No experience	152	48.1%
	Experience	164	51.9%
	Total	316	100.0%

Table 2b SQE1 FLK Assessment: Pilot Candidate Demographics			
Legal Education			
Variable	**Value**	**N**	**%**
Have you studied or are you studying the Legal Practice Course (LPC)?	No	70	22.2%
	Yes	246	77.8%
	Total	316	100.0%
Have you successfully completed the Graduate Diploma in Law ?	Not completed	254	80.4%
	Completed	62	19.6%
	Total	316	100.0%
Have you or will you have completed the compulsory modules of the LPC by the time of the pilot?	Not completed	79	25.0%
	Completed	237	75.0%
	Total	316	100.0%
Have you passed the LPC?	Not passed	260	82.3%
	Passed	56	17.7%
	Total	316	100.0%
If yes what was your overall grade/award?	Commendation	19	33.9%
	Distinction	20	35.7%
	Pass	17	30.4%
	Total*	56	100.0%
Do you have a UK university undergraduate degree in law ?	No	98	31.0%
	Yes	218	69.0%
	Total	316	100.0%
Do you have an undergraduate degree in law from a Russell Group University (E+W)?	No	245	77.5%
	Yes	71	22.5%
	Total	316	100.0%
Do you have an overseas University law degree?	No	291	92.1%
	Yes	25	7.9%
	Total	316	100.0%
What class of University law degree were you awarded?	Third	6	2.6%
	Two two	58	24.9%
	Two one	124	53.2%
	First	45	19.3%
	Total*	233	100.0%

* = variable not applicable to 'missing' candidates

Table 2c SQE1 FLK Assessment: Pilot Candidate Demographics			
Personal Demographics			
Variable	**Value**	**N**	**%**
Sex	Female	215	68.3%
	Male	100	31.7%
	Total*	315	100.0%
English 1st Language?	Yes	265	83.9%
	No	51	16.1%
	Total	316	100.0%
Religion classified	Christian	120	38.0%
	Muslim	37	11.7%
	None	119	37.7%
	Other (<10/religion) / Not stated	40	12.7%
	Total	316	100.0%
Sexual orientation	Bisexual	11	3.5%
	Gay/lesbian	17	5.4%
	Heterosexual/straight	275	87.0%
	Other / Not stated	13	4.1%
	Total	316	100.0%
Classified Ethnicity	Asian/Asian British	70	22.3%
	Black/Black British	37	11.8%
	Mixed/Multiple Ethnic Groups	16	5.1%
	Other Ethnic Group	7	2.1%
	White/White British	184	58.6%
	Total*	314	100.0%
Binary Ethnicity (Derived from the above)	BAME	130	41.4%
	White	184	58.6%
	Total*	314	100.0%
Gender ID different from birth?	No	297	94.3%
	Yes	18	5.7%
	Total*	315	100.0%
Age classified	Older =<1993	151	47.8%
	Younger >=1994	165	52.2%
	Total	316	100.0%
Disability as per Equality Act 2010?	No disability	292	93.3%
	Yes, a disability	21	6.7%
	Total*	313	100.0%
Reasonable adjustments requested?	No	298	94.3%
	Yes	18	5.7%
	Total	316	100.0%

* = A few candidates did not respond to item

How representative this group are of the ultimate candidature for the SQE is, of course, impossible to say at this point in time. A reason for the introduction of the new qualification system is to reduce barriers to entry into the profession. However, below we contrast key candidate demographics with those of the LPC candidature in 2016/17[9].

A number of aspects of the information in Table 2 are noteworthy:

Home background and work experience

- Just over half of candidates (55%) came from home backgrounds in which neither parent possessed a degree-level qualification
- Almost half (48%) of candidates had been educated at non-selective UK state schools

[9] https://www.sra.org.uk/sra/how-we-work/reports/authorisation-monitoring-activity-2016-17

- Just over half (52%) of candidates had practical legal experience (defined as experience either as an apprentice or trainee solicitor, CILEX, or paralegal)

Legal Education

- 77% of candidates had successfully completed a degree in law, with 90% of those at a UK university and 10% overseas
- The great majority of law graduates, 73%, reported a 2.1 or first-class degree
- 33% of the UK law graduates took their degree at a Russell Group University in England and Wales
- 18% of candidates had already passed the LPC with roughly a third each reporting distinction, commendation or pass; 75% of candidates had completed the compulsory modules at the time of the pilot exam
- 20% of candidates had successfully completed the Graduate Diploma in Law (GDL)

Personal demographics

- Females were represented more than twice as often as males, 68%:32%, paralleling the latest LPC data (2016/17) which shows a 64%:36% F:M split
- For the majority (84%), English was their first language
- 42% of candidates had BAME ethnicity (LPC 2016/7: 41%)
- 18 (6%) candidates had 'reasonable adjustments' and 21 (7%) (LPC 2016/7: 11.8%) regarded themselves as having a disability within the meaning of the Equality Act 2010

The Assessment

Using experience and methodology acquired in the development of the QLTS MCQ paper since 2011, three MCQ single best answer papers were developed according to the subject composition described in Table 1 (above). 316 candidates (see above) sat these three papers at a total of 44 Pearson VUE test centres, in the UK and overseas, over a two-day period in March 2019. Candidates provided demographic information at the time of registering for the test.

Candidates were randomised into two groups, such that half sat 60 of the questions in each of the three papers on day one, the other half sitting the same items on day two. Timing of the papers was such that candidates had 1.8 minutes per question.

Candidates' test scores were calculated for each of the three papers. Each paper's quality was assessed using routine measures of reliability (Cronbach's alpha) and precision (Standard error of measurement: SEm). The quality of test items was assessed using conventional indices of difficulty, discrimination (corrected item-total correlation), and response time. These analyses were supplemented by others based on Item Response Theory (IRT). Content review of all questions, informed by the statistical analysis, was conducted by solicitors. Candidates' overall scores on 360 items were then examined against individual demographics and, using multivariate analysis, to identify the key predictors.

Subject scores were correlated against paper scores. Data on test accuracy were calculated for the three 120-item tests, and, using generalisability theory in addition, then modelled for different test lengths and evaluated against current best assessment practice.

RESULTS

Results by Paper and Overall

Candidates' scores on the three papers are summarised in Table 3 and shown visually in the following histograms (Figures 1-3). Average candidate scores were approximately 50%.

Table 3: Candidate Scores by Paper

Paper	Number of Items	Minimum %	Maximum %	Mean %	SD %
Paper 1	120	25.00	85.00	50.70	11.69
Paper 2	120	17.50	72.50	46.81	11.34
Paper 3	120	28.33	80.00	51.94	10.13

Figure 1: Distribution of Paper 1 Score on 120 Items

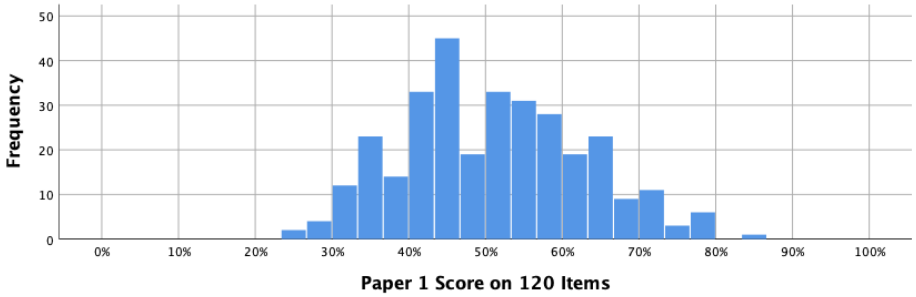


Figure 2: Distribution of Paper 2 Score on 120 Items

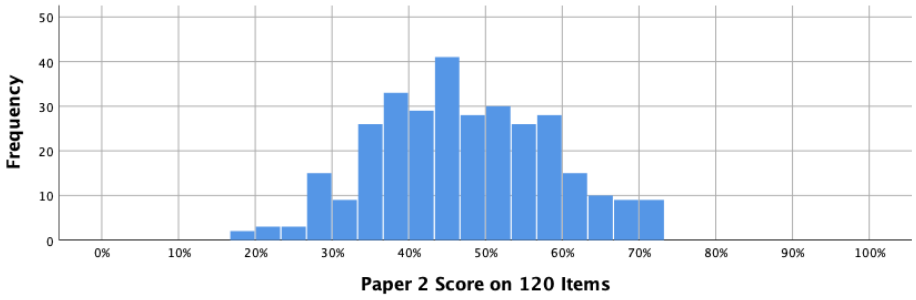
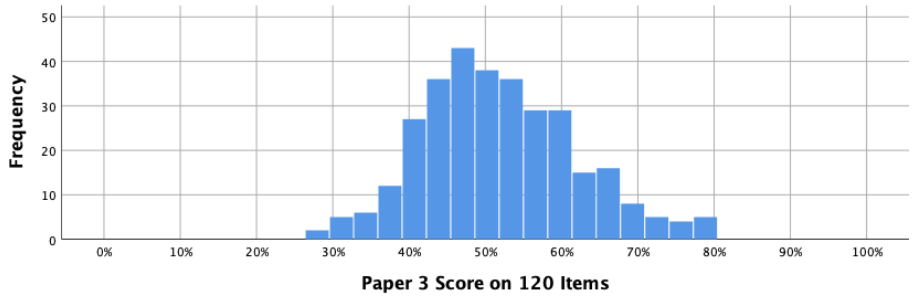


Figure 3: Distribution of Paper 3 Score on 120 Items



Test Statistics

For high stakes credentialing assessments, a reliability coefficient of 0.90 or better is desirable for a multiple choice test to support validating the outcomes. This is commonly seen by regulators as the

"gold standard" for a multiple choice test in high stakes licensing exams[10].

The reliability coefficient is also used to calculate a Standard Error of Measurement (SEm) which can be used to form a confidence interval around the observed score to show the likely range of scores on retesting with a paper covering similar content with different items. The SEm uses the reliability coefficient in its calculation, so as alpha values increase the SEm becomes smaller, and outcomes become more certain. For high stakes examinations, the reproducibility of pass/fail outcomes is critical, with low SEms desirable. Many psychometricians would recommend an SEm of no more than 4% for a high stakes multiple choice style credentialing exam.

Table 4 lists these quality statistics (alpha, a 95% confidence interval around alpha, and SEm) for each paper of 120 items. Each one nearly, but not quite, reached the key quality targets mentioned.

Table 4 Test Quality Statistics for three 120-item papers

Data	Paper 1	Paper 2	Paper 3
Candidates (n)	316	316	316
Items (n)	120	120	120
Cronbach's alpha (α)	0.88	0.87	0.84
95% CI for	0.86, 0.90	0.85, 0.89	0.81, 0.86
SEm (%)	4.05	4.08	4.02

Test Item Statistics

Table 5 lists the range, mean and standard deviation of the items for each of the three papers. 94.4% (340) of the 360 test items showed test difficulty between 0.15 and 0.85. While these thresholds are arbitrary, questions at the extreme ends are likely to contribute less to test quality and this would be considered in the review of the questions conducted by solicitors.

342 (95.0%) showed positive item discrimination. The point-biserial correlation between an item and the total test score if that item was excluded (corrected item-total correlation – CITC) provides an effective and straightforward measure of item discrimination. Higher positive correlations indicate that those doing well on the test overall were more likely to get the item correct, with a negative correlation indicating that lower scoring candidates overall were more likely to get the item right.

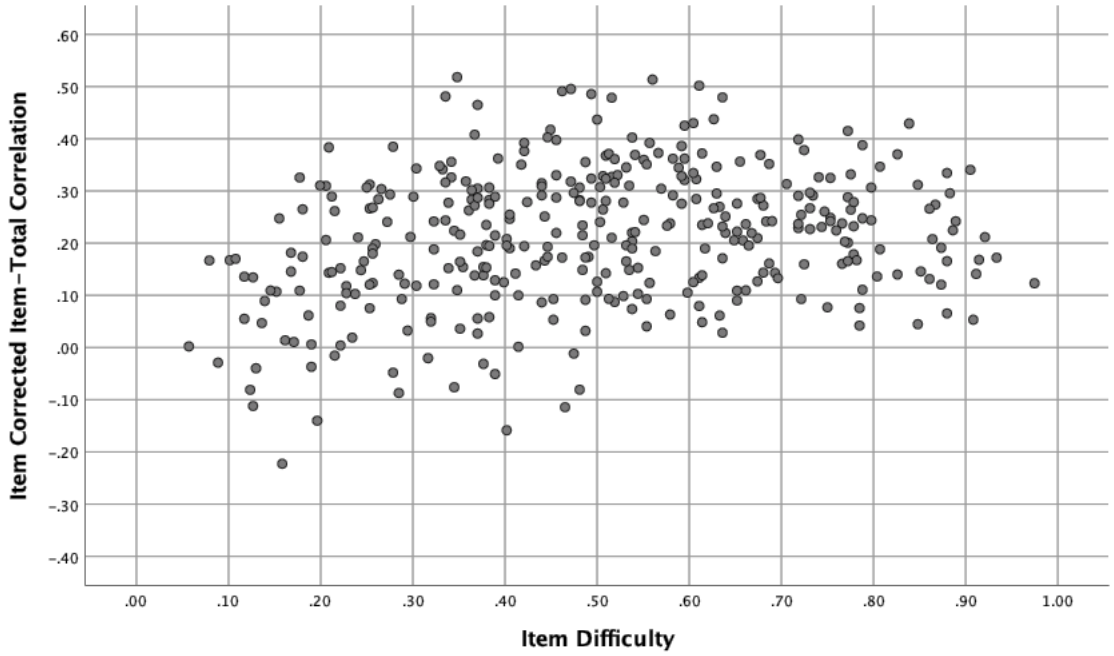
The average item response time was 87.5 seconds. This was well within the allocated time of 108 seconds per question and suggests that it may be possible to reduce this allocated time. However it should be borne in mind that candidates on the pilot were likely to take less time than those sitting in a high stakes 'real' exam. Average item response time per question ranged between 33 and 315 seconds, with only a few questions at the top end of the range. Just two items took candidates an average of longer than 3 minutes. Both of these items were reviewed by the academic team and are discussed below

[10] 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

Table 5: Range, mean and SD of Item Statistics						
Paper	**Number of Items**	**Test Statistic**	**Minimum**	**Maximum**	**Mean**	**SD**
Paper 1	120	Item Difficulty	0.11	0.93	0.51	0.21
		Corrected Item-Total Correlation	-0.16	0.50	0.23	0.13
		Median Item Time (s)	27.00	141.50	78.68	24.68
Paper 2	120	Item Difficulty	0.08	0.89	0.47	0.20
		Corrected Item-Total Correlation	-0.22	0.51	0.21	0.14
		Median Item Time (s)	30.50	255.50	81.21	25.99
Paper 3	120	Item Difficulty	0.06	0.97	0.52	0.22
		Corrected Item-Total Correlation	-0.11	0.52	0.19	0.11
		Median Item Time (s)	25.00	149.50	67.53	19.63

Figure 4 makes explicit the relationship between item difficulty and the corrected item-total correlation for each item. Each point in the scatterplot represents a single item.

Figure 4: Scatterplot of Item-Total Correlation against Item Difficulty (all 360 items)



A review of the response patterns for each item was carried out to identify items which had a distractor that was a more popular response than the correct answer. The response pattern is another factor considered by solicitors in their content review of questions. Of 360 items, the most popular answer for 270 of them was the correct option. Of the 90 items that contained a significant distractor, the correct option was the second most popular response for 51 of them, with a further 39 items having two or more responses more popular than the correct option.

Additional analyses using IRT

Alongside the Classical Test Theory (CTT) analyses (above), we also explored the value of IRT models in evaluating assessment quality. IRT examines the performance of items and candidates on the latent trait that an assessment is designed to measure, and both approaches provide distinct and useful information to support the validation process. However, we should note that the use of IRT for the pilot is essentially exploratory. Candidate numbers are unlikely to be sufficiently large in initial administrations to make the approach robust.

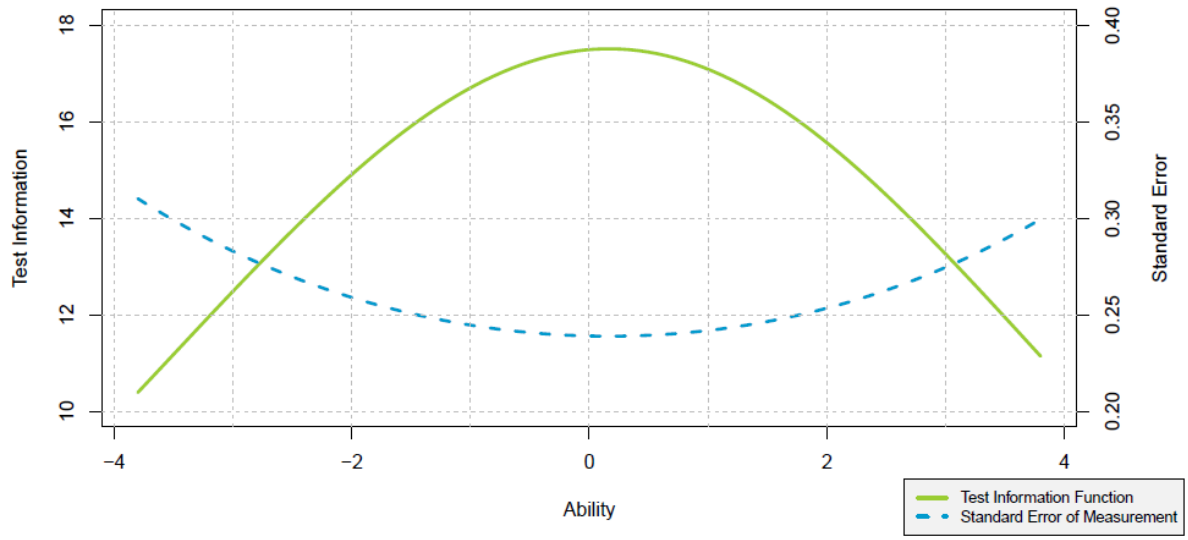
For the pilot, we focussed on the Rasch model and used all 360 items, treating them as a single assessment measuring a common latent trait to increase the reliability of the analysis and simplify

interpretation of findings. The choice of IRT model is dependent upon the sample size, so for a larger live exam, other IRT models may be viable.

Rasch modelling produces a range of statistics that are designed to help understand the nature of the test and the relationship between candidates and items. While it highlighted some items that performed at variance to the expected values, these were sometimes contradictory to those highlighted by CTT item statistics. As such, it demonstrated this analysis may add value but should not be relied upon as the sole method of reviewing items and refining the test.

IRT's Test Information Function (TIF) and corresponding SEm are shown across all 360 items in Figure 5, and this indicates that the assessment provided the most information and the least error for those in the middle of the ability range. Effectively, this demonstrates that the test provides the most information for those who are likely to be around the passing score, and that the assessment was not skewed to favour discrimination between low or high ability candidates. For a pass/fail examination this is very appropriate.

Figure 5: Test Information Function and SEm across candidate ability



Content Review by solicitor subject experts

All 360 items were analysed by solicitors in the light of their statistical performance on the test and, where applicable, their performance within previous QLTS tests. Due attention was paid to items which had been flagged by the various statistical indicators already mentioned, particularly where an item was flagged by multiple indicators. In general a negative discrimination index, measured here using the CITC, may be indicative of an unsatisfactory item and items which few candidates get correct contribute little to test reliability and precision. But subject-expert review may adduce considerations which outweigh the raw statistics' implications.

For example, one item which just 9% of candidates answered correctly (difficulty = 0.09) and was negatively discriminating (CITC = -0.03), tests a fundamental principle of land law concerning joint ownership of land. This would feature at an early stage in any programme of study, and practitioners and academics would regard it as simple and essential in understanding the ownership of land. Selection of any of the distractors represents a clear misunderstanding of the position at

law. Two other items with low discrimination coefficients (CITC = .053 and .056) related to key requirements of the statutory powers of advancement. These three questions showed otherwise high-scoring candidates' misconception of the position at law and examiners felt that the 'poor' discrimination needed to be outweighed by legal content review. Removing these items from the test would enhance the test quality statistics, but would impact on the test's blueprint coverage and validity.

The team also examined the few items which had apparently taken candidates a long time to answer. The question taking the longest time was on inheritance tax. Content experts (who were also experienced law teachers) took the view that with the appropriate knowledge and ability to apply it, it could be easily answered in less than a minute. Otherwise, a long time might be taken exploring different combinations of the numbers given in the question, hoping to reach one of the options as the answer. The other was again a tax question, and candidates who did not know any tax could have spent a long time getting to the wrong answer. These may have been questions candidates left unanswered before returning to them with their remaining time after completing the rest of the paper.

This review by content experts confirmed experience in the QLTS that statistical indicators should not be considered on their own and that questions should not be discounted or excluded solely on the basis of their statistical performance.

Overall, the review by content experts, taking into account the statistics, will be used to improve the question bank. The analysis suggests that there is modest room for improvement of individual items. No items were removed from the test as a result of the review, though a few items would likely be retired or rewritten on the basis of the information obtained in this pilot.

Item analysis: conclusions from the Candidate Survey

During the live test, candidate queries were recorded. Only two were made and both made legally incorrect assertions. A candidate feedback survey was carried out after the administration of the pilot with responses as follows (Table 6):

Table 6: Candidates' Responses to Questions about the Items

Question	N Respondents	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
"The FLK questions were clear"	201	15.4%	47.3%	19.4%	15.4%	2.5%
"The FLK questions covered appropriate knowledge"	200	9.5%	46.0%	15.5%	21.0%	8.0%
"The FLK questions were set at an appropriate level"	201	11.4%	39.8%	18.9%	22.9%	7.0%

Candidates' results according to separate Demographic Variables: 'univariate analyses'

Table 7 lists the background demographic variables mentioned earlier and summarises differential candidate performance across the whole 360-item examination according to the relevant sub-groups.

The table shows numbers, mean percentage score and standard deviation for each sub-group; then the ANOVA F-ratio and its significance level ("p"); an estimate of effect size (Cohen's d) for significant binary comparisons; and finally the group favoured by a significant outcome.

Considerable caution is needed in interpretation of these univariate statistics. Firstly, this is because of the small numbers of candidates involved, but secondly, because of substantially confounding relationships between individual variables.

In fact, these background variables are multiply confounded with implications as to how simple outcome differences on any variable are interpreted. For example, some 42% of candidates were of black, Asian and minority ethnic (BAME) ethnicity compared to 58% who were white. Overall 20% of candidates had taken the Graduate Diploma in Law (GDL). However, while 28% of white candidates had taken the GDL, only 8% of BAME candidates had done so (χ^2=18.26, df = 1, p <0.000).

Moreover, some of the significant differences may be seen as counter-intuitive; for instance that which shows that candidates with a UK university law degree perform worse than those without one. The sample of 316 candidates is comprised mainly of 60 people with a GDL and no UK law degree and 216 people with a law degree and no GDL. (A further 2 possess both and 38 possess neither.) The average score of the two groups is 58.1% (GDL only) and 48.5% (UK law degree only), highly significantly different (see Table 7; Anova F = 21.27 p < 0.000). The data do not show that possession of an undergraduate law degree is unhelpful towards performing well in SQE1, indeed its degree class predicts performance in the FLK (Table 7), but rather suggests that possession of the GDL is more helpful.

Similarly they do not show that passing the LPC is unhelpful towards performing well in SQE1. The group who had not passed the LPC included all those who were currently on the LPC. Similar factors apply in relation to work experience. Again the figures do not suggest that work experience is unhelpful, rather that confounding variables are at play.

Candidates' results according to multiple Demographic Variables: 'multivariate analyses'

In order to explore which are the best true predictors of candidates' total score within this sample, aside from individual differences in competence, multivariate analyses can be carried out that attend to multiple variables at the same time. In particular, with a sample of this size we can use multiple linear regression, with categorical variables such as ethnic group separated into individual "dummy" dichotomous variables such as 'No Religion – yes/no'. However even multivariate analysis cannot account for all factors and should be viewed with considerable caution especially in view of the relatively small numbers on the pilot, the complexity of the analysis, and the large number of variables.

Repeated explorations of the multiple available variables suggest that in this candidate group, the best and most significant predictors of candidates' final scores are as set out in the accompanying table (Table 8). But it should be recalled that all the demographics were self-reported and thus subject to inaccuracy. Moreover, the distribution of some variables included very small categories, making statistically significant results less likely with regard to them.

Table 7 SQE1 FLK Assessment: Differential Performance by Candidate Demographics (on all 360 items)

Area	Variable	Value	Mean %	N cands.	SD %	F	p	Significant difference or NS?	Effect size (for sig. binary comparisons): Cohen's d	Analysis favours?
A: Home background and work experience	Level of parental education when was 18	1+ degree level qualification	51.18	143	11.74	2.20	0.09	NS	-	-
		No formal qualifications	47.52	67	8.11					
		Not stated	46.71	7	10.52					
		Qualifications below degree level	49.62	99	9.09					
		Total	49.82	316	10.29					
	School type classified	Independent	53.47	33	12.87	2.09	0.08	NS	-	-
		Other / not stated	42.46	7	5.95					
		Outside UK	50.07	70	10.04					
		State: non-selective	49.57	151	9.63					
		State: selective	48.94	55	10.64					
		Total	49.82	316	10.29					
	Experience as CILEX, trainee or apprentice solicitor, or paralegal?	No experience	52.48	152	11.07	20.86	0.000	SIG	0.5 'medium'	No experience
		Experience	47.35	164	8.86					
		Total	49.82	316	10.29					
B: Legal Education	On or completed LPC?	No	43.97	70	8.86	31.85	0.000	SIG	0.7 'medium'	LPC on/completed
		Yes	51.48	246	10.08					
		Total	49.82	316	10.29					
	GDL successfully completed?	Not completed	47.89	254	9.43	53.11	0.000	SIG	1.0 'large'	GDL completed
		Completed	57.73	62	9.97					
		Total	49.82	316	10.29					
	Compulsory modules of the LPC completed?	Not completed	44.23	79	8.97	34.29	0.000	SIG	0.7 'medium'	Compulsory LPC modules completed
		Completed	51.68	237	10.04					
		Total	49.82	316	10.29					
	LPC Passed?	Not passed	50.58	260	10.72	8.14	0.005	SIG	0.4 'medium'	LPC not passed
		Passed	46.30	56	7.10					
		Total	49.82	316	10.29					
	LPC Outcome	Commendation	47.44	19	7.13	6.92	0.002	SIG	-	Distinction->Commendation->Pass
		Distinction	49.22	20	6.16					
		Pass	41.59	17	5.93					
		Total*	46.30	56	7.10					
	Do you have a UK university undergraduate degree in law?	No	52.71	98	11.76	11.62	0.001	SIG	0.4 'medium'	No UK university degree in law
		Yes	48.52	218	9.30					
		Total	49.82	316	10.29					
	Do you have an undergraduate degree in law from a Russell Group University (E+W)?	No	48.83	245	10.40	10.25	0.002	SIG	0.4 'medium'	Those with Russell group university degrees
		Yes	53.21	71	9.18					
		Total	49.82	316	10.29					
	Do you have an overseas university law degree?	No	50.34	291	10.23	9.86	0.002	SIG	0.6 'medium'	No o'seas law degree
		Yes	43.70	25	9.07					
		Total	49.82	316	10.29					
	Undergraduate degree class	Third	41.16	6	4.61	4.88	0.003	SIG	-	Those with better degree
		Two two	45.05	58	7.28					
		Two one	49.28	124	9.79					
		First	50.40	45	9.98					
		Total*	48.23	233	9.41					
C: Personal Demographics	Sex	Female	49.01	215	10.11	3.83	0.051	NS	-	-
		Male	51.43	100	10.50					
		Total*	49.78	315	10.28					
	English 1st Language?	Yes	50.35	265	10.15	4.44	0.036	SIG	0.3 'small'	English is 1st lang
		No	47.05	51	10.69					
		Total	49.82	316	10.29					
	Religion classified	Christian	48.95	120	10.44	10.44	0.000	SIG	-	No religion ->Christian ->Muslim
		Muslim	43.40	37	7.49					
		None	53.18	119	9.87					
		Other (<10/religion) / Not stated	48.37	40	9.92					
		Total	49.82	316	10.29					
	Sexual orientation	Bisexual	50.56	11	12.73	2.95	0.033	SIG	-	Other/not stated ->Gay/Lesbian ->Bisexual ->Heterosexual/ straight
		Gay/lesbian	50.82	17	4.94					
		Heterosexual/straight	49.35	275	10.24					
		Other / Not stated	57.84	13	11.91					
		Total	49.82	316	10.29					
	Classified Ethnicity	Asian/Asian British	46.02	70	9.68	9.67	0.000	SIG	-	Mixed ->White ->Asian ->Other ->Black
		Black/Black British	43.99	37	9.80					
		Mixed/Multiple Ethnic Groups	52.67	16	9.86					
		Other Ethnic Group	45.24	7	7.64					
		White/White British	52.31	184	9.85					
		Total*	49.79	314	10.30					
	Binary Ethnicity (Derived from the above)	BAME	46.22	130	9.88	29.12	0.000	SIG	0.6 'medium'	White candidates
		White	52.31	184	9.85					
		Total*	49.79	314	10.30					
	Gender ID different from birth?	No	49.64	297	10.26	0.909	0.341	NS	-	_
		Yes	52.02	18	10.78					
		Total*	49.78	315	10.28					
	Age classified	Older b->1993	49.12	151	10.55	1.33	0.250	NS	-	-
		Younger b1994->	50.46	165	10.04					
		Total	49.82	316	10.29					
	Disability as per Equality Act 2010?	No disasbility	49.65	292	10.43	0.509	0.476	NS	-	-
		Yes, a disability	51.31	21	8.08					
		Total*	49.76	313	10.28					
	Reasonable adjustments requested?	No	49.96	298	10.30	0.942	0.332	NS	-	-
		Yes	47.53	18	10.14					
		Total	49.82	316	10.29					

* A few candidates made no response to these items

Table 8 SQE1 FLK Assessment: summary of multivariate analysis				
Variable	**R Square Change**	**F Change**	**Sig. F Change**	**Approx. Score Variance Explained**
GDL successfully completed	0.148	54.34	p < 0.000	15%
Russell Group University (E&W) Law Degree	0.081	32.94	p < 0.000	8%
No Religion	0.050	21.40	p < 0.000	5%
No Practical Legal Experience*	0.042	19.09	p < 0.000	4%
Male Gender	0.020	9.37	p < 0.002	2%
White Binary Ethnicity	0.015	7.07	p < 0.008	2%

* as CILEX, Trainee Solicitor, Paralegal, Solicitor Apprentice

Thus, very substantial predictive value appears to be provided by two educational factors: having successfully completed the GDL accounted for 15% of score variance, while possessing an English or Welsh Russell Group university law degree accounted for a further 8% of the score variance.

Univariate and multivariate analyses: Equality issues

Demographic and equality analyses of the candidates on the pilot must be viewed with considerable caution given the small size of the sample, confounding variables, the fact that characteristics were self-declared, and, for the multivariate analysis, the complexity of the statistical model used. However initial tentative indications are that educational factors are key predicators of success in the FLK exam. While this provides evidence for the concurrent validity of the exam, it is also a cause for concern given the confounding of these educational factors with membership of minority groups protected under the Equality Act 2010. Kaplan will continue to work with the SRA to ensure that protected groups are not unfairly disadvantaged while maintaining the standards of the assessment.

Setting a passing standard

For this pilot delivery of the FLK papers, it was not considered appropriate to 'pass' or 'fail' the volunteer candidates who were provided with their own scores and the overall distributions of scores.

We did however investigate various standard methods of setting the pass mark for FLK which can be used alone or in conjunction, including Angoff's and Hofstee's[11]. The Angoff approach involves a panel of trained judges providing an estimate of the proportion of minimally competent candidates who would answer each question correctly. The sum of all these estimates is used to create a cut score aligned to the minimally competent (borderline) standard for an assessment. To this borderline standard, in high stakes examinations an amount is normally added, depending upon test precision, to accommodate test unreliability. Hofstee's method requires judges to indicate an acceptable range for the potential passing score, and an acceptable range for the pass rate, against which is plotted the cumulative score distribution. The standard is set to where a line drawn between these parameters intersects the observed cumulative data.

No final decisions were made about the pass mark of the papers on the pilot but it seems likely that it would be above 50%. On the published sample FLK questions, the SRA have indicated that the pass mark would be in the region of 55 – 60%[12].

[11] See for instance Zieky MJ, Perie M, Livingston SA. Cutscores: *A Manual for Setting Standards of Performance on Educational and Occupational Tests.* Educational Testing Service; 2008.

[12] https://www.sra.org.uk/sra/policy/sqe/sqe1-functioning-legal-knowledge-assessment-specification/sample-questions/

FURTHER DISCUSSION: Paper length and issues of compensation

Generalizability Coefficients and Decision Study

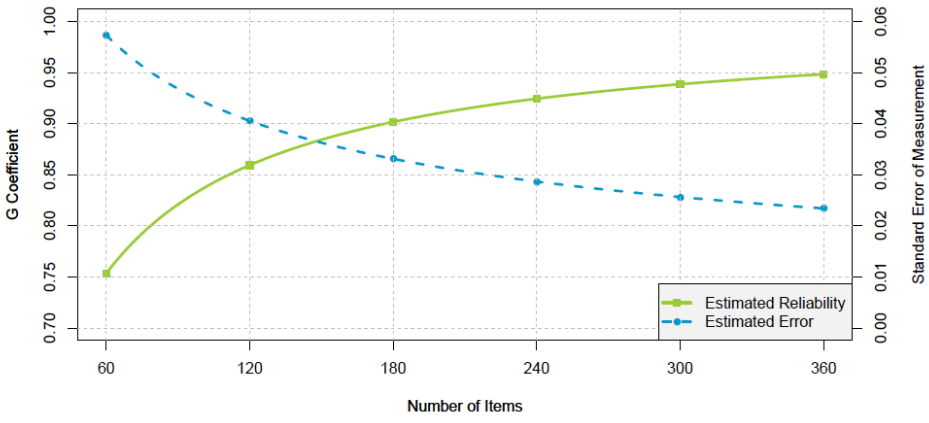
Overall the pilot FLK performed well. However, three papers of 120 questions did not quite reach key quality targets for a credentialing exam (see Table 4). Changes to the design were, therefore, considered with the aid of Generalizability (G) Theory[13].

G-theory quantifies the contribution of different sources of variance associated with test scores and provides a value for reliability and its associated SEm based on the test design and outcomes. For this purpose, an analysis of variance was performed to decompose the variation in observed scores into variance components for candidates, items on each of the papers, and the residual (error) variance. For a standard MCT assessment, a simple G-theory model will produce the same values as CTT analysis.

Using G-theory to examine reliability has distinct advantage over other CTT methods. While alpha and the G coefficient relate to the observed performance, G-theory can use the estimated variance components to model changes to the assessment design as part of a D-Study. Using G-theory values as a starting point, a D-study can predict reliability and precision for various test conditions. Specifically, because our equations used for G-theory analysis include values for the number of items and tests, we can manipulate these values to find the optimal reliability and SEm for an assessment using the observed variance.

The G coefficients and corresponding SEm are presented in Figure 6, which shows the impact of more/less items on the generalizability (reliability) and SEm of scores. The more sophisticated G-study gives a similar conclusion to our earlier observations around test reliability. Tests of 120 items should be close to, but may never reach, levels of reliability and precision expected as a minimum for high stakes assessments such as a national licensing qualification. However, tests of 180 questions are predicted to have a reliability coefficient of 0.90 and SEm of 3.31%, thus meeting the threshold for the acceptable psychometric standard to provide reproducible evaluations of competency of candidates taking the assessment.

Figure 6: Test Reliability and Measurement Error by Test Length



[13] G Theory was originally described in: Cronbach LJ, Nageswari R & Gleser GC (1963). Theory of generalizability: a liberation of reliability theory. The British Journal of Statistical Psychology, 16, 137-163. The more recent generally-used reference is: Brennan RL (2001). Generalizability Theory. New York: Springer-Verlag.

Test length: how many Papers?

As shown above generalizability analysis suggests that two papers of 180 items will reach acceptable levels of reliability and accuracy.

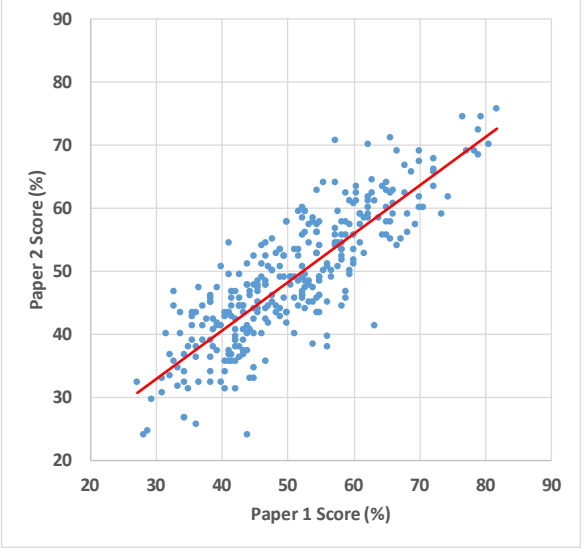
We therefore looked at the statistics from the pilot and considered the actual performance of the pilot questions as two papers divided as follows: Paper 1 – Business Law and Practice; Dispute Resolution; Contract; Tort; Public and Administrative Law, the Legal System of England and Wales; Paper 2 – Property Practice; Land Law, Wills and the Administration of Estates and Trusts, Solicitors Accounts; Criminal Law and Practice.

Analysis of the reliability of the two sets of 180 items (papers as above) shows improved alpha values of 0.91 and 0.90 respectively, for the 180-item test forms. Table 9 lists these quality statistics. Divided into two papers of 180 items each, the 95% confidence intervals show a range between 0.88 and 0.92. The SEm for each also drops to 3.3% as certainty in the outcomes increases due to the longer test length and increased reliability. Figure 7 shows the correlation between candidates' marks on two 180-item papers ($r = 0.84$, $p < 0.0001$) indicating a high level of test-retest reliability and concurrent validity across both.

Table 9 Modelling Test Quality Data for two 180-item papers

Data	Paper 1	Paper 2
Candidates (n)	316	316
Items (n)	180	180
Cronbach's alpha (α)	0.91	0.90
95% CI for	0.90, 0.92	0.88, 0.91
SEm (%)	3.3	3.3

Figure 7: Scatterplot showing Correlation of Marks between two 180-item papers



It is however worth noting that due to the nature of the pilot and candidate pool, we might expect some inflation of reliability in the pilot due to a wider range of scores than might be recorded during a summative assessment.

Compensation Issues

A key concern of stakeholders with using either three papers of 120 items or two papers of 180 items each is compensation between subjects. Will candidates be able to pass by selectively choosing certain areas to focus on and avoiding other areas completely? Experience from other professional qualification exams suggests that this is unlikely (and it would certainly be a very high risk strategy). However, we examined the pilot statistics to look at the extent to which compensation was taking place.

The following figures (Figure 8 and Figure 9) provide a visual depiction 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.

One (very thin) horizontal line depicts each candidate's score on the relevant test. Those with the top scores were coloured dark green, those in the next quintile were coloured light green, next blue, 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 dark green band; below that would be a solid light green 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 light green and blue 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, particularly near quintile boundaries) across subjects, as candidates will have spent more time preparing for the exam and rely less on their existing knowledge.

While these statistics are reassuring, this is a matter which should be kept under review following the introduction of the SQE.

Figure 8: Compensation within 3 x 120 question tests

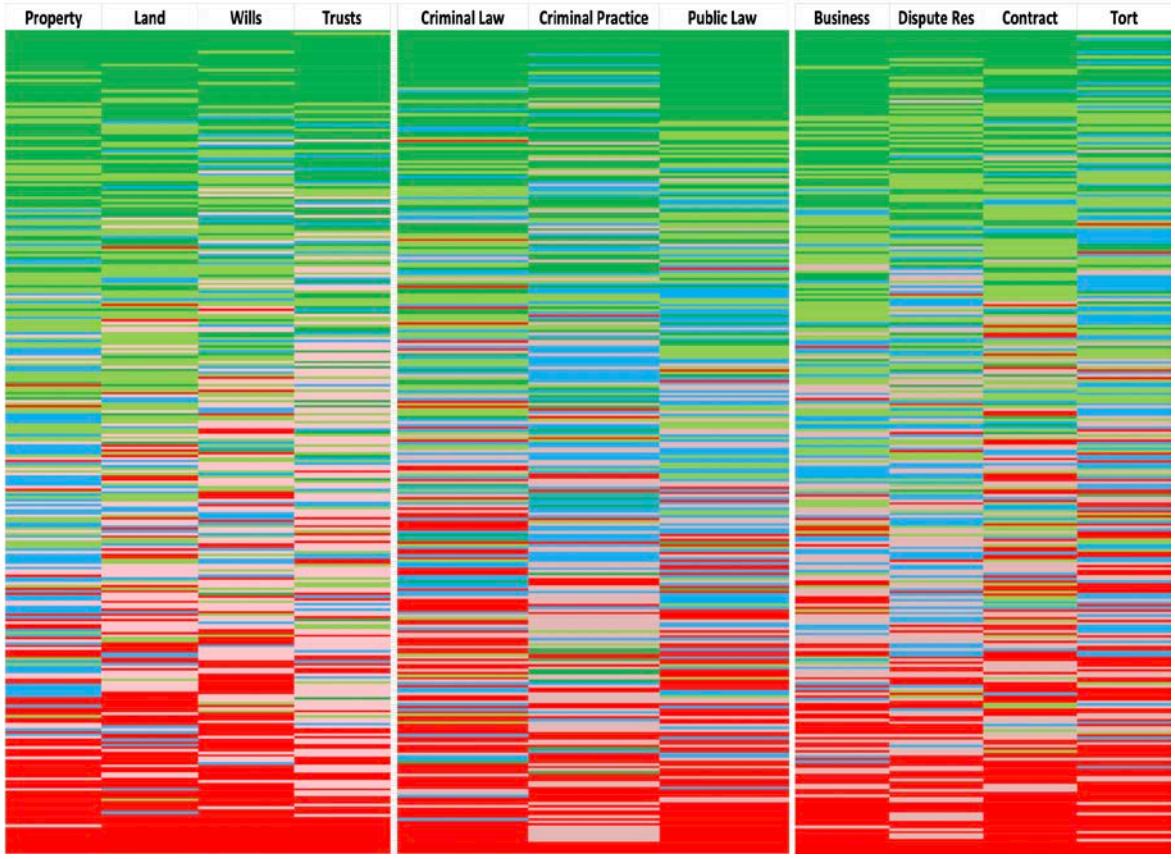
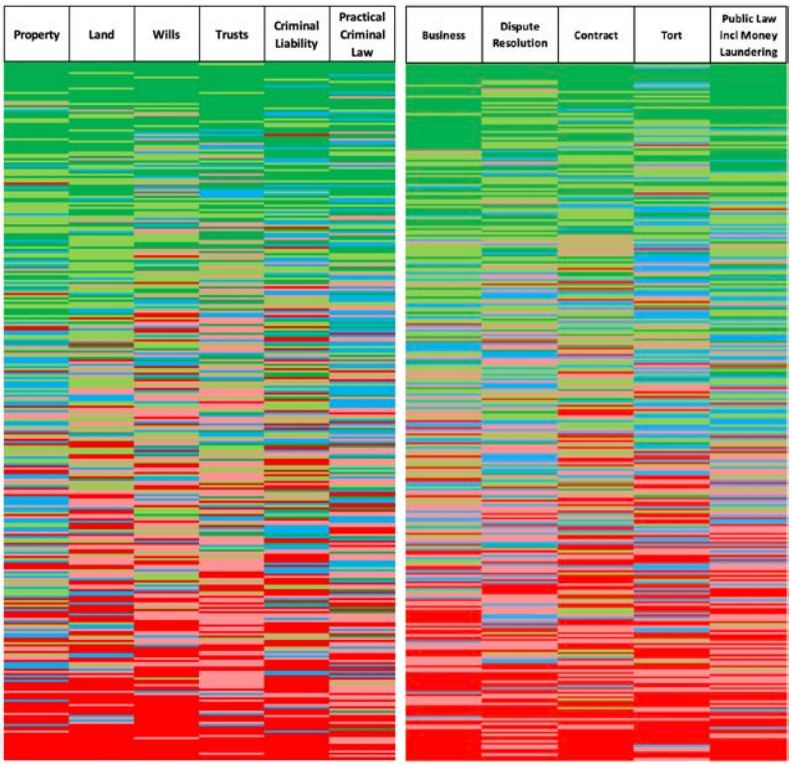


Figure 9: Compensation within 2 x 180 question tests



Conclusion

This paper has considered statistical and psychometric issues arising from the SQE1 pilot delivery of the FLK. Overall, we regard the SQE1 FLK pilot as a successful delivery providing the statistical and psychometric evidence necessary, alongside other factors, for decision making about the SQE assessment design.

Suitable candidates were selected who were as representative as possible of the group who would take the SQE both as to prior education and demographics. The pilot questions produced a good range of marks from 17% - 85 % with the average at around 50%. The pilot will give a clear basis for decisions about the design, content and performance of 3 x 120 item papers, but has also considered proposed alternative test forms, and in particular 2 x 180 items.

All 360 questions used in the pilot were reviewed using a variety of statistical methods and all were then reviewed by the Academic Team in the light of these statistics with particular emphasis on questions which were flagged by multiple statistical indicators. The review reinforced the conclusion that while statistical data is informative in analysis of items, expert legal analysis of the content is equally important. Poorly performing items may reveal unexpected gaps in candidates' understanding of basic concepts, rather than poor drafting. The questions were validated by the candidate survey. Overall, our conclusion was that while there was modest room for improvement of some items in the question bank, this was unlikely to make a significant difference to the test quality statistics as a whole.

Demographic and equality analyses of the candidates in this pilot must be viewed with considerable caution given the small size of the sample, confounding of demographic variables, the fact that characteristics were self-declared, and, for the multivariate analysis, the complexity of the statistical model used. However, background educational factors seemed to be those which predicted most score variance. Completion of the GDL and a Russell Group University law degree seemed to be the most significant. While this provides support for the concurrent validity of the exam, it is also a cause for concern given the confounding of these educational factors with membership of minority groups protected under the Equality Act 2010. Kaplan will continue to work with the SRA to ensure that protected groups are not unfairly disadvantaged while maintaining the standards of the assessment.

Reliability (reproducibility) and accuracy of outcomes (precision) are key to a high stakes licensing exam. Three exams of 120 items each, very nearly, but not quite, reached the levels of reliability and precision commonly regarded as the "gold standard" in national licensing exams. Two exams of 180 items each did reach the levels of reliability and precision commonly considered desirable by regulators. Questions of compensation between different subject areas which are a key concern of stakeholders were reviewed for each of these designs and showed that while there was some compensation between subjects it was limited.

March 2020.