Annex E

BAR STANDARDS BOARD

REGULATING BARRISTERS

Aptitude Test Consultation

Contents

Appendices

Information on the format of the proposed test and example questions can be found at http://www.barstandardsboard.org.uk/media/1344440/watsonglaser_form_ab_example_questions.pdf

Responses to the consultation Questions (page 39) should be submitted by
29 February 2012 to:

Aptitude Test Consultation
Bar Standards Board
289-293 High Holborn
London WC1V 7HZ

Or by email to: AptitudeTest@BarStandardsBoard.org.uk

Responses are welcome from all those who may have views or evidence relating to issues raised in this paper. It would be helpful; if responses could be fully presented with detailed reasons given for comments, as well as any underpinning evidence.

The BSB will summarise the responses received and will normally publish responses on its website. If you do not wish your response to be published, please make that clear in your response.

1 Background

1.1 Proposed alteration

It is proposed that, in addition to existing entry requirements as specified in the Bar Training Regulations (BTRs[1]), applicants for the postgraduate Bar Professional Training Course (BPTC), formerly known as the Bar Vocational Course (BVC), be required to attain a minimum pass threshold on the Bar Course Aptitude Test[2] (BCAT), which has been carefully developed and piloted specifically for this purpose.

It is proposed that this change should commence with the cohort of candidates applying from November 2012 to start the course in September 2013.

It is proposed that those taking the test will be allowed an unlimited number re sits. Candidates with adequate skills but one (or more) unsuccessful attempts at the BCAT will not be prevented from undertaking the BPTC once they have achieved the requisite score.

It is proposed that the test will be run by Pearson Vue and therefore available to be taken at any Pearson Vue testing centre, of which there are hundreds worldwide.

1.2 Background, rationale and problems with the current system

Due to pressure from the OFT, the monopoly of the Inns of Court School of Law in delivering the Bar Course ended and, following a validation exercise, eight institutions in England and Wales were approved to deliver the course. The period 1997-2006 saw a steady increase in student numbers, a widening gap between the number of graduates and pupillages available and a high rate of failure on the course. Concerns about the standards on the Bar Vocational Course and the mismatch between the growing number of BVC graduates and reducing numbers of Pupillages had increased significantly by the time of the establishment of the Bar Standards Board. A major Review of the BVC was therefore commissioned by the BSB with a Working Group Chaired by Derek Wood QC, which produced recommendations after working on the project between October 2007 and July 2008.

Due to concerns expressed at the standard and apparent unsuitability of many students on the course, the BSB adopted the recommendation (amongst others) of the Wood Review that standards should be raised on entry and exit to the Bar Course, which was renamed the BPTC. This included the recommendation that the entry requirements for the course be amended to include an Aptitude Test that would be fair to all applicants and test the requisite skills (rather than, for example, excluding some suitable candidates by raising academic entry requirements). It was therefore proposed that, in addition to existing requirements as specified by the Consolidated Regulations[3], such a test should be developed and that this test should be compulsory for all students to pass before they are able to start a Bar Course (the new BPTC).

[1] Can be found at http://www.barstandardsboard.org.uk/media/1344498/bartrainingregulations-1092011.pdf
[2] The Aptitude Test is based on the established and recognised Watson Glaser Critical Thinking Test which is used by some law firms in recruitment assessment days and by the Graduate Management Admissions Council. This is discussed in more detail later in this report.
[3] Superseded by the Bar Training Regulations(BTRs) from 1 September 2009

This recommendation was made in order to address concerns about the abilities of students undertaking the course. These concerns arose not only from a significant body of anecdotal evidence, but also the experience of panel members who had visited the Bar Vocational Course Providers and from speaking to interest groups as discussed later in this report.

A threshold requirement or pass level should be identified to ensure that only suitable candidates should be admitted to the course. Any student unable to achieve the pass threshold in the test would not be eligible to start the course but an unlimited number of re sits would be permitted so that candidates subsequently attaining the required threshold would then be admitted.

A detailed programme of development and implementation then followed:

October 2007 – July 2008	Wood Review including consultation
September 2008 – July 2009	Development of specification for the test, tendering process for a suitable provider and independent external consultant/statistician
September 2009 – July 2010	First pilot (c 200 BVC students took the test to determine basic viability, approach to assessment and validity as a method of identifying suitable candidates for the BPTC)
September 2010 – July 2011	Second pilot (c 1500 BPTC students took the test to establish its validity, test the assessments themselves and determine the pass threshold or ‘cut score’)
July 2011 – November 2011	Analysis of findings and recommendations
December 2011	BSB formal consideration
December 2011 – February 2012	Further consultation period

It was determined that the test would be run by Pearson Vue, available to be taken at any of their testing centres, of which there are over 150 in the UK and others in 165 countries worldwide. The BSB also duly engaged an independent consultant who was employed to assess whether the chosen test was appropriate for use and also contracted with Pearson Vue to run two pilots of the test, as detailed above.

1.2.1 Standards and Failure rates on the Bar Course (BVC)

The Working Group reported that the existing entry requirements had been found to be insufficient in their present form to maintain the necessary standards on entry for the Bar Course: “The student body includes graduates who are so far lacking in the qualities needed for successful practice at the Bar.... that they would never obtain pupillage, however many pupillages were available.” This is borne out by statistical evidence of low standards on the course. In fact, for the academic years 2003-09 only an average of 64% of BVC students passed all modules on the first attempt:

AY	First time pass rate
2003-4	69%
2004-5	63%
2005-6	61%
2006-7	60%
2007-8	65%
2008-9	65%

It is also particularly important to note that after two re sits[4] approximately 10% of students still did not manage to pass the course. This demonstrates that students are admitted who are not capable of passing the course. There is of course the cost implication to consider for students who are admitted to the course. Fees for the Bar Course are constantly rising, with most being between about £10,000 and £16,000. Add to this the cost of living and also the potential cost of a year of one's life spent on a course which, if ultimately failed, results in no professional qualification or academic award; the BSB believes that there is a duty to ensure that only those who have a hope of passing are admitted.

In addition, the course is highly skills-based and interactive, with typically over 70% of the course being delivered by small group practical sessions. Students are paired up regularly for role plays and are separated into groups during other interactive sessions. Students who do not have the aptitude for the course inevitably find scenarios difficult to follow or are unable to participate in groups and thus slow the entire session down. This has a cumulative and considerable effect not only on the rest of that group but also the tutor and the resources of the course as a whole. That is to say that a proportion of course failures are likely to be due to basic lack of aptitude and others' marks are potentially brought down by this interaction.

1.2.2 Research and consultation

A survey was conducted on all students on the course in 2008 as part of the consultation in relation to the review of the BVC. A good response was obtained from over 500 current and recent BVC students in 2008 who were asked for their views on the course, across all Providers and with a good mix of gender, academic and ethnic backgrounds amongst respondents. Students were asked about their views on a range of subjects including curriculum, teaching, assessments, resources etc, as well as their overall experience. 49% of respondents stated that their experience on the course was adversely affected by the learning needs of other students.

Much discussion therefore took place with stakeholders concerning the proposal for an aptitude test, including practitioners, teaching staff, students and consumer groups.

[4] On the BVC two re sits were allowed, this has now been reduced to one.

Comments were noted in over 30 focus groups and other meetings. A summary of these comments can be found in the appendices.

1.2.3 Conclusions of the Working Group

The evidence provided above and appended clearly demonstrates why secure entry requirements need to be determined and adhered to, and why the current entry requirements need to be strengthened. There is a need for the Regulator to set and monitor this as a specific entry requirement. This is essential as a regulatory activity; a system which is regulated rather than left to individual Providers of the Bar Course will ensure fairness and consistency for all students.

Responsible for regulation and Quality Assurance of the Bar Professional Training Course, the BSB is concerned to ensure that students are gaining a valuable experience on the course, and that the students exiting the Course are of a high enough standard to begin a pupillage at the Bar of England and Wales, should they wish to do so. By ensuring that only those who are capable of passing the course are eligible, the BSB will be doing more to ensure that the experience of other students is not affected, due to the highly interactive nature of the course, by a possible lower level of aptitude in peers, and equally that individual students who do not have the requisite skills for the Course are not recklessly permitted to attend the Course and pay the large fees and expenses associated with it. The Bar Standards Board takes its obligations to ensure equality of entry to the BPTC very seriously.

Section 4 of the Legal Services Act states that: “the Board must assist in the maintenance and development of standards in relation to......the education and training of persons so authorised.” For the reason stated above, introducing this additional entry requirement would help the Board in doing so by improving the educational experience, improving the standards on the course and ensuring that only those who have a real prospect of passing the course actually undertake it.

1.2.4 The chosen test

The Working Group made an additional recommendation that the test should have the following characteristics:

> “(1) It must test two skills separately: analytical and critical reasoning and fluency in the English language. Candidates must pass both parts.
> (2) It must be taken by all prospective BVC *(now BPTC)* students irrespective of their background.
> (3) It must be available to anyone who wishes to take it at any stage in their career after entry into university.
> (4) Candidates should be able to take the test any number of times until the pass mark is reached.
> (5) The test must be set at least twice a year.
> (6) It must be an on-line test capable of being taken at a number of centres within and outside the United Kingdom.
> (7) The test must be capable of being objectively marked.

(8) The cost of taking the test must be met by the candidate. It must therefore be inexpensive.

(9) There will be no interviews.[5]"

A tendering process was undertaken during 2008 looking at tests and test providers available. An invitation to tender was sent to seven companies and several responses were considered carefully. Pearson Vue was identified as the most suitable provider to use and the Watson Glaser Critical Thinking test was chosen as best to satisfy the criteria above[6].

1.3 Aptitude and IELTS

In 2008 the Working Group recommended that:

"....part of the test could be built upon the IELTS, which BVC students from overseas must already undertake and achieve a standard of 7.5. However the experience of some of our members shows that a 7.5 score in IELTS does not indicate the level of ability which we think is necessary. The BSB should, in our view, engage a consultant to advise on the correct format and level of both elements of the test; and it could if it wished run it as a voluntary pilot test for the BVC intake for 2009."

The two tests are designed with very distinct aims. The IELTS test descriptor for score band 7 (BSB requires 7.5) is to identify "A Good user; someone who has operational command of the language, though with occasional inaccuracies, inappropriacies and misunderstandings in some situations. Generally handles complex language well and understands detailed reasoning." The Aptitude Test is "designed to assess critical thinking in five areas: inference, recognition of assumptions, deduction, interpretation and evaluation of arguments."

Whilst it remains the ultimate aim of the BSB to use one test to assess English language and aptitude for the course, it has been confirmed by our independent consultant that the Watson Glaser Critical Thinking test is not designed to test English language. It would not be fair to say that someone with a good enough command of the English language to pass an IELTS test will also possess the requisite aptitude to complete the course. Likewise it would not be fair to say that someone with the ability to pass an Aptitude Test will have a good enough grasp of the English language (e.g. in speaking or listening) to complete the course satisfactorily. This is why the BSB considers there is a clear need for both tests as entry requirements for the course, until such a time as a unified test can be developed.

Furthermore, the English language requirement has encountered difficulties in its application. The BSB applied to change the BTRs[7] to require proof of English language skills for all applicants of the BPTC; however it was advised that the submission should be withdrawn, considering it to be disproportionate. However, EU legislation indicates that there is a problem in applying the rule selectively because of the EU Directive 2005/36/EC on recognition of professional qualifications and freedom of movement of persons/professionals

[5] Bar Standards Board Wood Report 2008

[6] The Aptitude Test is based on the established and recognised Watson Glaser Test which is used by some law firms in recruitment assessment days and by the Graduate Management Admissions Council. This is discussed in more detail later in this report.

[7] Can be found at http://www.barstandardsboard.org.uk/media/1344498/bartrainingregulations-1092011.pdf

in the EU. The English language rule has therefore recently been amended so that it applies to all candidates but requires that:

“Applicants should be able to demonstrate that their oral and written English language ability is at least equivalent to:

a. a minimum score of 7.5 in each section of the IELTS academic test;

b. a minimum score of 28 in each part of the internet based TOEFL test; or

c. a minimum score of 73 in each part of the Pearson Test of English (academic)

On entry to the course students will be required to sign a statement that they are aware that this standard is required of all students who enter the BPTC, and that they consider that they have met it. Those with any doubt as to the level of their English skills, are strongly advised to undertake one of the above tests before enrolling on the course.”

This is clearly a difficult rule to apply and for students to work with, and although the tests are separate, it demonstrates the real need for the Aptitude Test to be introduced so that at least students with the requisite aptitude are on the course, and its performance can be sufficiently monitored while a test for dual purpose can be developed.

1.4 Possible adverse effects

As with any testing mechanism, there may be instances where very capable students do not score the requisite pass threshold. The statistics of this are analysed later in this document and also in the Report from the independent consultant appended, but it must be remembered that the test will come with an unlimited number of re sits. This will have the effect that a capable student who does not score the requisite score due to illness or any other reason will be able to re sit the test at their own pace (but before starting the course) and will then have the opportunity to show their eligibility for the course on another occasion.

One aspect that must be looked at is the possibility of disproportionate effect on any particular demographic group. Again, statistics found from the research will be looked at later in this report. There is a chance that any marked differences in the scores achieved in the BCAT will be mirrored in performance in the course, with the result that this adverse effect is in fact proportionate and a legitimate means to the aim.

There is, of course, a cost implication associated with the Aptitude Test, which may cause some adverse impact to those who cannot afford the cost. Judging from the fact that the current fees for the course range from £16,095 to £10,300[8] and that the cost of the BCAT is £67 which represents approximately 0.6% of the cost of the very cheapest course (and an even smaller proportion where fees are higher than this). The cost is slightly higher for international students due to higher administration costs, but this will still represent a cost of less than 1% of the course fees. The cost of undertaking the course is naturally compounded

[8] Taken from the Course Contact Information document available on the BSB Website

by the need to meet travel and living expenses etc as well. The Bar Standards Board considers that an increase in costs by this low percentage is justified by the necessity for the additional requirement. The additional requirement would result in positive effects on the courses and on the educational experience of the students. As mentioned earlier in this document, the course is of a highly interactive and practical nature and thus the Aptitude of the students undertaking it is of paramount importance in relation to the experience of the rest of the student body and the best and most efficient use of resources at the BPTC Providers. Evidence of the impact of weak candidates on the learning experience of others has been demonstrated in annual visits to courses, including meetings with students and staff and in teaching observations. This was also confirmed by students according to feedback questionnaires during the BVC Review, as mentioned above.

The introduction of the BCAT is not intended to reduce numbers of students studying the BPTC; it is intended to ensure that only candidates who are likely to pass the course are actually eligible to undertake it. It must be recognised that a consequence *may* be that numbers are reduced at first, but the rule change is certainly not looking to encourage this. The rule change aims at raising exit standards on the course by improving the educational experience and ensuring that up to £16,000 (not including living costs) is not taken irresponsibly from students who really have no realistic prospect of passing the course.

1.5 Aim

Having regard to the rationale above, the BSB’s aim in relation to this rule change is to ensure that only those with a realistic chance of passing are allowed to start the Bar Course. This will save some weak students up to £16,000 on course fees and improve the educational experience, due to the interactive nature of the course, for those who are successful in gaining a place.

2 The First Pilot January – August 2010

The first pilot was undertaken in early 2010. 182 students participated in this pilot ("the First Pilot").

2.1 Results of first Pilot

The results of the First Pilot showed a good correlation between marks on the BCAT and final grade on the course, as shown by the following table:

BPTC Result	Average Aptitude Test score
Outstanding	67.6
Very Competent	61.2
Competent	50.7
Referred	43.2

The relationship between the test and examination scores can also be expressed as a correlation. This is a number between -1 and +1 which reflects the strength of a relationship; the higher the absolute value, the stronger the relationship. For selection purposes correlations of 0.3 and above are desirable but those above 0.2 are still useful. The correlation between the BCAT and examination scores in the First Pilot was 0.62 showing a very strong correlation.

2.2 Need for a further pilot

There were, however, a number of factors which meant that this pilot could not be considered to be sufficiently robust to enable immediate full implementation of the test:

1. Those who took the pilot test were self-selecting and therefore did not constitute a representative sample of all students on the course. In particular weaker students were likely to be under-represented.
2. The test was taken very late in the academic year whereas it would normally be taken before any exposure to the course.
3. The numbers undertaking the test represented a very low percentage of students on the course and were too low to allow full evaluation of the test and determination of a pass score.
4. There was some evidence of differences in performance levels for members of some demographic groups which required further investigation.

For these reasons the BSB undertook to hold a second, larger pilot ("the Second Pilot).

3 The Second Pilot November 2010 – November 2011

3.1 Hypothesis

Following the initial consultation and recommendation, tendering period, discussion with the independent consultant and potential provider, and the first pilot, the purpose of the second pilot was to test the hypothesis that:

The proposed Aptitude Test would be an effective additional means of assessing candidates' suitability for the course, able to reliably predict the likelihood of a candidate's success or failure on the course. It would fulfil the aim of ensuring the right standards in delivery of the course, for protection of consumers and that students would not needlessly waste time and money on a course that would not ultimately prove beneficial to them in terms of their future career aims (to assist individuals).

3.2 Method

The BSB sought advice from the independent consultant, an independent statistician and the Bar Council's Research Manager, in relation to the minimum sample necessary for a further pilot (the Second Pilot) to:

(a) assure the validity of the test itself using a larger cohort,
(b) assure the validity of the questions proposed to be used and
(c) set a pass mark threshold.

On that advice, the Second Pilot was aimed to test all students who commenced the course in September 2010; the tests were conducted in November 2010. Although there were around 1,700 students on the course, 1,501 took the Aptitude Test. Later the test scores were matched with the results obtained on the Bar Course.

The Providers were asked to report the grade (competent, outstanding etc.) and final examination score for students after the first examination sittings in summer 2011 and the second sittings in late summer, early autumn. From the first sittings grade data was supplied for 1370 individuals and a specific final examination score for 728 people. Following the re sits, grade data was provided for a further 298 students, and an additional 396 received a revised grade. Examination scores were provided for 570 students following re sits including 256 who had examination scores assigned at first sit. 88 existing examination results were revised following re sits by 1 point or more.

The Independent Consultant reported to the BSB on the results in three stages. All three stages of this Report are included as Appendices to this document.

The first stage of the Report consisted of analysing monitoring data and looking at the scores on the Aptitude Test according to the monitoring data given by students when they took the test. Just fewer than 1000 students completed the demographic information requested. The following figures provide a summary of the demographic make-up of the sample:

- 55% Female

- 56% White
- 92% English Primary Language
- 66% Aged under 25
- 9% Have a disability

Monitoring data was collected in order to check whether the test had any unintended and unjustifiable consequences for any particular demographic. Further to this, A Level results, Degree class and Degree institution information was collected, in order to investigate whether existing information might provide predictions as accurately as the BCAT. Fewer of the sample responded to these questions, and the responses were not as reliable, but fortunately there was sufficient data to perform some analysis.

Although it would be desirable, no data was collected regarding socio-economic status. This was due to the difficulty in formulating a short meaningful measure and the desire to minimise the demographic information requested to increase response rates. While there was a chance that collection of this data might show that the existence of the BCAT was capable of widening access further than the entry requirements and selection criteria currently in operation do, since only candidates who had successfully obtained a place on the course through the current entry arrangements were tested as part of this pilot, such questions were unlikely to return a meaningful conclusion. This question should be investigated if and when the test is operational.

The Second Stage Report looked at the relationship between the BCAT scores, A levels and Degree class and the results from the first sit scores of Bar Course examinations and compared their ability to predict success on the course. This stage could not analyse the test's ability to predict failure as students who had failed the course were not at this time confirmed.

The third Stage of the Report looked at the re sit (final sit available on the BPTC) scores from the Bar Course compared with BCAT scores. This stage, although looking at fewer variables, was of optimum importance as it would produce results to show the test's ability to predict those who were likely to fail the course and provide information to support determination of a pass threshold. Those students who failed at this stage had failed the BPTC, save for any extenuating circumstances which may exist.

An interesting report written by Dr Chris Dewberry[9] identified 12 requirements for effectively evaluating psychometric tests. These requirements are stated below, together with comments relating to the BCAT:

1. The purpose of the test should be clarified.

The purpose of the test is to identify those who will or will not pass the BPTC as determined by the BVC Review.

2. Establishing content validity.

[9] Aptitude Testing and the Legal Profession; 6 June 2011. This was commissioned by the LSB and looked at and commented on by the BSB's Independent Consultant. The full comments can be found in the appendices.

A job analysis was conducted in 2009 to establish content and test validity for its purpose and the test was chosen to match the identified cognitive skills profile.

3. Consideration of a range of different techniques.

Since the test is designed to identify weak cognitive skills, clearly a cognitive test is most appropriate.

4. Evidence of construct validity.

This is not necessary in this context.

5. Reliability.

The accuracy of the test was evaluated in the Third Stage Report and the reliability will be above 0.8 for the proposed test length.

6. Criterion for criterion related validity.

The clear criterion for the test is performance on the BPTC which is measured by outcome grade and examination score.

7. Criterion related validity.

The correlation found in the Stage Three Report was 0.51. This is highly statistically significant and compares well with the standard of 0.3 cited in Dewberry's report.

8. Incremental validity.

The design of the pilot studies shows incremental validity since they were performed on a pre-selected group of candidates who were subject to selection criteria to get onto the course. Incremental validity has also been shown of the BCAT over educational qualifications.

9. Subgroup differences.

Subgroup differences with respect to gender, primary language, age, ethnic group and disability have been examined. Data was not available for social class for the reasons discussed on the previous page.

10. Practice opportunities.

Example questions for the test are already openly available. The test Provider can be asked to provide a full length practice test once the Aptitude Test is operational.

11. Selection decision rules.

Any decision on the cut score is supported by the evidence of its potential effects in Stage 3 of the Report. It is based on identifying candidates who are not likely to pass the course.

12. Regular reporting.

The Bar Standards Board is committed to operating continual monitoring of the implementation of this test and its effects; these plans are addressed in more detail towards the end of this report.

The conclusion from reading Dewberry's paper is that the requirements set out in the report have either already been addressed in the programme of work or are due to be addressed before the test goes live. Issues still to be addressed include setting the cut score for the test, defining the ongoing reporting procedures and specifying how practice opportunities will be provided (all of these are addressed in Stages 2 and 3 of the report).

3.3 Score Distribution

For a test to be effective it should differentiate well between high and low scorers, which means that a range of scores is required. For the second pilot test students completed one of 5 versions of the test to enable the piloting of a large pool of questions.

Figure 1 of the Report Stage 1 (in the Appendices) shows the distribution of scores on each of the 5 versions; scores are expressed as percentage correct. ***It can be seen that there is a good range of scores. There are some very low scores that would allow the use of the test to select out poor students.***

In interpreting these results it is important to remember that lower scores in this sample may be due to poor motivation among students who completed the test, rather than poor ability. The recorded time spent on the test is an indicator of this. For these students there was no important outcome depending on performance on the test so they may not have performed at their best. On the other hand it is also likely that there will be more people of low ability among applicants than these students who have passed the selection criteria to gain their place on the course.

3.4 Predicting Performance

3.4.1 Relationship between test scores and course outcomes

A good relationship between test scores and course outcomes is critical to using the test to identify people unlikely to pass the course.

A correlation is a statistic which provides an estimate of the size of the relationship between two variables. A correlation ranges from -1 to 1 where 1 is a perfect linear relationship and -1 is a perfect negative relationship (i.e. as one variable rises the other falls). In this case a positive relationship between examination results and test scores is desirable. A correlation of 0.3 or above is desirable in using a test for selection although even values of 0.2 can be useful. If a finding is statistically significant it means it is unlikely to be due to chance.

Section 3.1 of the Stage Three Report shows the correlation between final examination results and test score for the 988 students for whom examination scores are available was 0.51. This value is highly statistically significant and shows that there is a strong relationship between the test and examination performance. The 99% confidence interval ranges from 0.45 to 0.57. This means that it can be concluded with a high degree of probability that there is a correlation of not less than 0.45 between the test and exam scores.

According to further calculations, the Report explains that the BPTC Examination score is predicted to rise by about half a standard deviation for every standard deviation rise in the BCAT score.

3.4.2 Predicting BPTC Grades

Table 3.3 of Stage Three Report shows that 77% of those who were graded Outstanding were in the top two quintiles (score bands) on the test whereas 70% of those who failed and 64% of those who ended up being referred after re sits were in the bottom two score bands. There were 109 failing students following the re sit results and 132 were referred for further examinations. Figure 3.1 of the Report stage 3 shows that failed students and those referred after re sits tend to score equally low on the BCAT whereas the majority of those graded outstanding have high scores.

The relationship is not perfect. There is a small percentage of those graded Outstanding or Very Competent that are in the lowest two bands of the test. Some of these may be students who did not invest any effort in completing the test and therefore have scores which do not reflect their ability***.*** Of course a similar effect would be true of any selection rule. Like those with low test scores, they will be exceptions to the general rule that those with better qualifications tend to do better on the course. For example, there will be some people with third class Degrees who might have done well on the course. ***Allowing applicants to re sit the test multiple times ensures that the test does not prevent a talented candidate from showing their eligibility for the course due to a single poor test score.***

As well as its ability to predict candidates likely to fail, it may be worth considering the additional effect that the BCAT could have in assisting the Bar Course Providers in selecting candidates. Although the test is intended to identify those who are likely to fail the course and therefore ensure that only capable candidates are eligible to undertake the BPTC, this report shows clearly that the test is also capable of identifying candidates who are likely to perform better on the course as well. This means that if BCAT scores are shared with Providers, with the candidates' consent, it could improve their ability to make informed choices about the candidates they select. This would help to fully inform the Provider and the student of the likely abilities of that student on the Bar Course and give confidence to all those who complete the Bar course and indeed increase confidence in those who have completed the Course.

3.4.3 Alternative Predictors

In the current data set there is information about A Level (or equivalent) results and first Degree grades that could provide similar information to BCAT scores. This information is only available for part of the sample. The difference statistic in the tables in the Reports in the Appendices (see definition in the Reports) provides an estimate of the size of the difference with values below 0.3 categorised as small, around 0.5 in the mid range and above 0.8 as high.

3.4.3.1 Scores by A Level Results

Secondary school results were available for some of the sample and these were translated into UCAS points where possible.

UCAS points in this pilot may be unreliable because

- Some participants may not have provided a complete record and therefore their points allocation may underestimate their performance.
- Some participants reported only up to 200 points but this would be insufficient for university entrance without alternative qualifications.

There is therefore some inconsistency in the results for the lowest scoring groups but Table 7 of the Stage 1 Report shows ***there is a strong relationship between the test scores and A Level results, despite these problems with the data.*** There is a difference of over 0.8 standard deviations which is considered high, showing a strong relationship.

3.4.3.2 Scores by Degree Class

Only a small proportion of the sample responded to the question regarding their undergraduate degree class. ***As with other educational results there is a strong relationship; a difference statistic of over 0.8 is shown between BCAT scores for those with a first and a lower second.*** Figure 8 of the Stage 1 Report clearly shows that those with first class Degrees performed considerably better on the BCAT.

3.4.3.3 Scores by Undergraduate Institution

Table 9 of the Stage One Report shows the results broken down by first Degree institution for those institutions with at least 5 respondents. There is a range of results with students from Oxford having the highest overall score on the Test and those from newer universities tending to have lower average scores. ***Table 6.7 in the Stage Two Report shows a moderate difference in BCAT scores for those from Russell Group universities compared with other institutions.*** This is likely to be related to the greater selectivity of the Russell group universities in choosing students initially. In fact, in this data set those who had attended Russell Group institutions had an average of 90 more A Level points than those who had attended other universities. Although this could therefore be used as a fairly effective indicator of performance, it would be inappropriate to be effectively counting on an entirely different institution to assist and guide a course's selection criteria/entry requirements.

3.4.4 Comparison of BCAT and Educational Qualifications as Predictors

Table 4.1 of the Stage Three Report shows correlations of 0.19, 0.46 and 0.31 between examination scores and A level points, degree class and whether the student attended a Russell Group university. These correlations are all statistically significant but lower than the BCAT correlation of 0.52.

Table 4.2 of the Stage Three Report looks at similar information to Table 4.1 but evaluates whether the test can add any predictive power after A Levels and Degree Class have been taken into account. Further details on how this was done can be found in the Report in the Appendices to this document.

The table therefore shows that:

- A Levels predict outcome on the BPTC,

- the BCAT predicts better than A Levels alone – 25% more variance explained,
- Degree results predict outcome on the BPTC,
- the BCAT predicts better than Degree alone – 14% more variance explained, and
- ***furthermore, the BCAT predicts better than A Levels and Degree information combined – 9% more variance explain.***

While degree class seems a potential alternative predictor from these results, it is already used as a requirement for entry on the course with only those with at least a 2.2 (or equivalent) eligible. This could be increased to a 2.1 or above but degree class does not allow any finer distinctions to be made. 30% of UK honours degrees awarded in 2009/10 were graded 2.2[10]

3.4.5 Summary of prediction results

The BCAT is a very strong predictor of BPTC examination results.
It is better than all educational qualifications combined as a predictor. Degree class is also a good predictor but the degree class scale is not finely differentiated which makes unsuitable to use further as a course entry qualification.

3.5 Demographic Group Comparisons

Comparisons were made according to gender, primary language, age, ethnic origin and disability. Most results are presented in a table and some with a box plot figure. Full representations are made in the reports in the appendices to this document. The difference statistic in the tables in the Reports in the Appendices (see definition in the Reports) provides an estimate of the size of the difference with values below 0.3 categorised as small, around 0.5 in the mid range and above 0.8 as high.

Where differences were found in BCAT scores for different demographic groups and sample size allowed, regression analysis was undertaken to see whether these differences reflected performance differences in course outcomes or whether they were an artefact of the test.

3.5.1 Scores by Gender

Table 2 of the Report Stage 1 shows a small difference (less than 0.3) in scores between men and women; ***with men scoring marginally higher than women. The difference is small but it does reach statistical significance.***

The regression analysis including gender showed no evidence of any bias with respect to gender in predicting course outcomes with BCAT.

3.5.2 Scores by Primary Language

Figure 3 of the Report Stage One shows that those with English as a primary language perform a little better than those for whom English is a second or later language. ***The difference in test scores is a little larger than for gender but still in the small range at***

[10] Source UK Higher Education Statistics Agency

less than 0.3. Table 6.1 of the Stage 3 report shows a larger difference in examination results. Section 6.2 of the Stage 3 report finds similar and significant correlations between BCAT and examination results for both the Primary and non-Primary language groups but the sample was too small for regression analysis to be performed. ***However there is no evidence of bias against those for whom English is not a primary language from the existing results.***

3.5.3 Scores by Age

Table 4 of the Stage One Report shows that there is some decrease in BCAT scores with age and the difference is largest between those in their late 20s and those in their 40s. ***The differences overall do not reach statistical significance.*** The sample size is quite small for the older groups. The largest d (difference) statistic is at the top of the small effect range, between 0.3 and 0.5.

Table 6.2 of the Stage Three Report details the regression analysis. There is a statistically significant but small direct impact of age. ***Although the effect reaches statistical significance it is small enough to have little or no meaningful impact.***

3.5.4 Scores by Ethnic Group

Table 5 of the Stage 1 Report shows larger differences by ethnic group with three quarters of a standard deviation difference in test scores between White candidates and others. Figure 5 of the Report shows that although there are substantial average differences there are high as well as low scorers in each group. Table 5 shows that similar differences are evident in Degree Class.

Examination results were available for 348 students who described themselves as white and 270 from other ethnic groups including 173 from various Asian groups. Only the Asian group was large enough to consider separately in the regression analysis.

Table 6.3 of the Report Stage 3 shows that the BCAT would tend, if anything, to over predict the performance of Asian students and therefore although they tend to score lower on the Test, there is no bias against them as these scores are reflected in lower Course grades and examination scores for this group.

Table 6.4 of the Report Stage 3 shows similar results for the comparison of white students with those of any other (non-Asian) ethnic background. There is a significant impact of ethnic origin on BCAT scores but this does not result in discrimination against the lower performing group.

3.5.5 Scores by Disability

75 students reported that they had some kind of disability. These results should be treated with caution as the numbers in the disabled group are quite small and therefore are likely to be subject to some variability.

Table 6.5 of the Stage Three Report shows that overall those with a disability have a similar level of performance on the test than those without. The largest group is those with dyslexia and their performance is on average better than that for the group that did not have any disability. Similar and significant correlations were found between BCAT and examination results for both the disabled and non-disabled groups but the sample was too small for regression analysis to be performed. ***These results suggest there is no prima facie evidence that candidates with disabilities would be disadvantaged by having to take the test.***

It should be noted that, in the trial, accommodations were not offered for students with disabilities. Such students would be allowed accommodations such as additional time etc. when the test is in operational use. This is further discussed later in this document.

4 Discussion Points

4.1 Cut score

Varying the cut score can reduce or increase the failure rate on the course. Being more aggressive with the cut score will reduce failure rate on the course, but it might mean that there are more people who might have passed the course who are not eligible due to failing the BCAT. Keeping the cut score low would mean that fewer people are prevented from undertaking the course but it could reduce the proportionality of the introduction of a test at all as it may only identify a very few of those who are likely to fail.

There are also two factors to be borne in mind when setting the cut score:

- applicants will be allowed to re sit the BCAT, and
- equality and diversity considerations and the risks of disproportionally excluding members of certain groups.

The aim of introducing the test is to prevent students that do not have the necessary ability to complete the course effectively from taking the course. Therefore Section 5 of the Stage 3 Report explores the impact of a variety of test cut scores in reducing the failure and referral rate and increasing the pass rate. It should be remembered that this sample is pre-selected using existing selection processes so the impact is incremental over current methods of selecting candidates.

4.1.1 Determining cut scores

Other approaches to determining a cut score are discussed in section 5 of the Stage Three Report. These concentrate on the boundary between Competent and Very Competent since although Competent is a passing grade it is a marginal one and to have a good chance of success a student should have the potential to achieve Very Competent so that if their performance falls below their potential they can still pass the course. This includes use of prediction equations and looking at the lowest scores obtained by students at each grade and the 5th percentile scores at each grade. These suggest cut scores between -2.56 and -0.53.

4.1.2 Impact of cut scores on selection and success rates

Table 5.1 of the Report Stage Three shows the impact of applying different cut scores to the current sample. Part time students have been excluded from this analysis since it is unclear whether these students will successfully complete the course. This leaves a total of 1234 students with identified grades. Going from the bottom to the top of the table, as the cut score becomes lower, it is less selective.

A cut score of -0.5 (high)

Of those who fail the examination or are referred again after re sits, less than half would have passed the test with this cut score. Those who received a marginal pass on the course (graded Competent) had a 60% chance of passing the test. In contrast 90% of those who obtained an Outstanding grade and 83% of those who were graded Very Competent would have passed the test. However, more than 15% of students that passed the course with a grade of Very Competent or Outstanding are ‘false negatives’ - they achieve highly on the course but would have failed the test. This figure may be somewhat exaggerated due to some students not taking the test as seriously as they would have, had their eligibility for the course depended on it. Further, there was no opportunity to retake for students who underperformed as there would be for real applicants. ***Nevertheless, this suggests that a cut score of -0.5 would be too high.***

A cut score of -0.75 (high)

This still results in a 50% reduction in the number of failures and over 43% reduction in referrals. The number of false negatives, those who achieve a high grade on the course despite failing the test is a little lower at 14% overall. At only 1% lower than using -0.5 cut score***, this still may represent too many false negatives and therefore be too high a cut score.***

Intermediate cut scores (-1.34, -1.25 and -1)

The impact of intermediate cut scores would have resulted in 21%, 16% and 10% of the students being ineligible to take the course according to results in the pilot. All three reduce the failure and referral rate substantially although not as much as the higher cut scores. They have a lower impact on those with the best course outcomes with 5%, 9% and 13% false negatives for a cut score of -1.34, -1.25 and -1 respectively. ***These cut scores seem to provide a better balance of reducing failures without excluding people who might have passed the course.***

Cut score of -1.5 (low)

Only 3% of students would have failed this cut score so it has very little impact on the pass rate generally although 11% of people who failed the course would have failed the test. Considering that about 10% of students fail the course outright, this would represent a tiny proportion, ***so would be too low for the introduction of the Aptitude Test to have any real effect.***

Tables 5.2 to 5.7 of the Stage Three Report in the Appendices to this document show very clearly lists of how each cut score would affect candidates which allows a detailed comparison of the different cut scores.

Table 5.8 shows the effect of various cut scores on not just course failures but also first sit pass rates. The highest of the three scores looked at (-1) reduces the failures at first sitting by 40% and the large number of referrals by 30%. ***There is a bigger impact on those referred who go on to fail. Just over half these students would have passed the test. In contrast over 80% of those who were referred but go on to pass the course well would have passed the test with the highest of these cut scores.***

4.1.3 Impact of cut scores on demographic groups

Section 7 of the Report Stage 3 looks at the effect of different cut scores on various demographic groups, focusing on the mid range cut scores -1.34, -1.25 and -1. A key statistic is the relative selection ratio. This shows the relative likelihood of passing the test for a member of one group relative to a comparison group. For example a value of 0.95 means that members of the identified group have a probability of passing the test that is 95% of that of the comparison group. Where the relative selection ratio is lower than 0.8 then the selection process breaks the ‘four-fifths’ rule and is considered to have a significant adverse impact. ***Gender, primary language, age and disability all show very minor differences in pass rates for all three cut scores and the selection ratios are consistently above 0.9.***

For the ethnic group comparisons where the largest differences in test scores were seen, all of the selection ratio comparisons are in the target region of 0.8 or above, however the results for the highest cut score (-1) for the Asian group are on the border of that target zone. ***A cut score of -1.25 would give a ratio sitting comfortably above 0.8, so within acceptable levels.***

4.1.4 Cut score summary

Different methods of identifying a suitable cut score have been looked at, the results of which are:

- Generally, scores between -1.25 and -1 provided a marked reduction in students who go on to fail the course without creating an enormous barrier for applicants or excluding many students who had good course outcomes. It must also be borne in mind that unlimited number of re sits would reduce this effect further.
- Identifying a cut score which will predict a desired examination result suggests a cut score of up to -0.53.
- Direct inspection of the test scores associated with each grade outcome suggests a cut score of -1.34.
- Cut scores of -1.25 and below show minimal levels of adverse impact even for the demographic groups with the larges score differences

As this is the introduction of a new test, it would be advisable to begin with a cut score at the lower end of the optimum range. The reason for this is firstly that the test is in any case not designed or intended to select out the best students and so it is appropriate to keep the cut score low. Secondly, the BSB endeavours to continually monitor the test's performance and the option will of course remain to raise the cut score for a more significant effect later. ***For that reason we have been advised that a cut score of between -1.34 and -1.25 is desirable.***

4.2 Alternatives

The BCAT could be used in an advisory manner rather than specifying a minimum score as an entry requirement. This could work in two ways:

- Advisory to students. Students would be able to take the test in order to assess their own aptitude for the BPTC. Score reports would include BCAT scores and predicted performance on the BPTC expressed as an outcome grade or a percentile with respect to the full cohort of students on the course. This may result in some students of weak standard choosing not to pursue undertaking the course. In reality it would not be likely to deter many students, indeed nor would many be likely to even opt to take the BCAT. This measure would therefore be unlikely to have any great effect on the entry standards and thus it would be unlikely to make any substantial improvement to the course experience.

- Advisory to Providers. Students would be required to take the test and their score would be shared with Providers to inform the decision making during the application process, in addition to information already collected in the application. The combination would undoubtedly improve Providers' ability to predict which students are likely to be successful on the course, but it must be borne in mind that the test is not intended to be used for narrowing access with another selection test. It is intended to reduce numbers of weak students on the course and thus to reduce fail rates on the course.

This report has identified in earlier paragraphs why the current language rule is not satisfactory as an alternative to using a BCAT. The language tests which are accepted by the BSB do not test a candidate's aptitude for the course in terms of reasoning skills and merely look at their language skills in isolation. Moreover, difficulties in application of the English language rule mean that it cannot be relied on to have much effect on the standards and experience of the course.

Some Providers of the BPTC have explored the option of using interviews as part of their selection process to ensure that they recruit only candidates who they can be confident are suitable for the course and have a prospect of gaining a pupillage. This has been successfully operated in one small Provider but carries with it the following problems:

- The large proportion of international applicants makes the logistics of interviews extremely difficult. This is made worse by visa issues.
- Interview sessions or days use a large amount of resources in terms of accommodation, time, materials and expenses, particularly when candidates may need to travel long distances for the meetings.

- Providers with larger cohorts of up to 350 students would find it impossible to find the time to conduct interviews for all candidates before making offers.

Selective use of the BCAT would entail the same problems with application as the English language requirement and would be very likely to receive challenge as being discriminatory on whichever group it was decided to test.

As part of the Review of the Bar Course, extensive and detailed consideration was given to the development and implementation of an Aptitude Test, in particular in relation to other ways of improving standards that could be considered. Limitation of numbers on the BVC was not possible, or favourable, due to competition law issues. It would require the approval of the Lord Chancellor/Secretary of State under the procedures set out in Schedule 4 of CLSA 1990 (as amended). Limitation of numbers was also opposed by the Advisory Committee on Legal Education and Conduct (ACLEC).

Another alternative, which was carefully considered during the BVC Review of 2008, was to raise the entry requirement to an upper second class Degree (2:1), however after discussion and consultation this was regarded as inappropriate due to inconsistencies in awards across the UK and overseas.

Finally, this report, in the rationale section, has identified why continuing with the entry requirements as they stand is not considered an option by the BSB.

5 The Regulatory Objectives and the Better Regulation Principles

5.1 The Regulatory objectives

5.1.1 Protecting and promoting the public interest

The number of undergraduate, postgraduate and training places in law, and the number of employment opportunities, are demand driven. Individuals should be free to pursue a career in law but with the knowledge that it is a highly competitive area with limited places available in firms, chambers, employed practice and government. Candidates should therefore be made aware of the high standards of training for the Bar of England and Wales and steps are also necessary to ensure that only suitable candidates undertake training.

The need for candidates to possess adequate skills on entry to training is clear, due in particular to the interactive nature of training on the course. The possession of not only academic knowledge, but also appropriate critical reasoning, use of language and other skills, is fundamental to the concept of providing high quality legal services in the public interest. Study of the BPTC demands a high level of ability and the public interest is best met using a specified entry requirement, applied fairly to all applicants. Moreover, it is in the public interest that the learning experience at Bar School should be of the highest quality and not adversely affected by weaker students during small group sessions, group discussions or while working as pairs in advocacy skills sessions.

Although the existing entry requirements (a Qualifying Law Degree or a Qualifying Degree followed by successful completion of a Conversion Course (BTR 18) awarded at first or second class honours), go a long way to ensuring a minimum entry level, the standards of Degrees are not always easy to assess and vary considerably between institutions both at home and overseas. It is thus essential for an additional measure to be put in place which is fair to all candidates. While it is true that some students do eventually practise in overseas jurisdictions, the training is for the Bar of England and Wales and all must reach the necessary standard for this. It is no longer the case that a Degree from a UK university will always be set at the required standard for entry to postgraduate legal training. In addition, a Degree normally tests academic proficiency rather than aptitude or skills.

5.1.2 Supporting the constitutional principles of the Rule of Law

Implementation of robust requirements for entry to the Bar Course can only serve to help uphold the principles of the Rule of Law, in helping to ensure the quality of the training courses and the standards of students.

5.1.3 Improving access to justice

The Bar Standards Board acknowledges that a requirement for all candidates to take the BCAT is likely to have only a limited impact on the objective of improving access of the public to justice. However the proposed regime will promote improvement of access to justice by helping to ensure that the students who exit the Bar Professional Training Course have experienced the highest standard of training and themselves are of a requisite standard. Maintaining high standards on the Bar Course will inevitably feed through into the Bar, resulting in many talented and capable Barristers available to represent the public.

5.1.4 Protecting and promoting the interests of consumers

In order to protect and promote the interests of those who use the services of the Bar, action must be taken to ensure that consumers can make informed choices about quality, access and value. It is thus essential to consider entry requirements for the BPTC (inputs) and not only outputs (ie passing the course). It is unrealistic to suggest that outputs alone can ensure the quality of practitioners, due to the interactive nature of the course. Training and the student experience must be of the highest level in order to ensure the highest level of outputs. It is certain that consumers feel more comfortable employing a service when they are assured of the high standards of training.

It is also important to note that potential students are also consumers, albeit to the Bar Course rather than to the Bar. The cost of the course is so high now and will only continue to rise in the future that it is truly irresponsible for the regulators not to ensure that only those likely to pass the course actually end up parting with such large amounts of money, not to mention the cost of living alongside undertaking a Bar Course.

5.1.5 Promoting competition in the provision of services

As has been addressed several times throughout this submission, adding an entry requirement in the form of an entry Aptitude Test will doubtless increase public confidence and standards at the Bar. In turn, this can only help to promote competition between highly trained practitioners.

The idea of competition between Providers must also be considered. The rising cost of the course and apparent discrepancies between the standards of students who are currently accepted onto different Bar Courses would be well addresses by the introduction of a uniform entry test ensuring that all students are of a minimum standard, not only in academic achievement (as ensured by the current entry regulations) but also in aptitude. As has been mentioned earlier in this report, it seems that standards of Degree, even across the UK, can vary considerably.

5.1.6 Encouraging an independent, strong, diverse and effective legal profession

The Bar Standards Board does not consider that using an Aptitude Entry Test will have any effect; adverse or otherwise, on the independence of the legal profession.

The BSB is committed to promoting diversity in the profession so that those with the right abilities are able to make a career as a barrister irrespective of their background, race, religion, gender, sexual orientation, disability or age. This means that all candidates must be treated fairly, in the same way, as they apply for the BPTC. The application of the BCAT for all candidates will not involve any additional implications for those with disabilities. Reasonable adjustments are made by the provider (Pearson Vue) at all their test centres, as appropriate.

The concept of an open and fair system, applicable in the same way to all, underlies the recommendation of the Wood Report[11] to introduce an Aptitude test for all for admission to the Bar Course. The pilots have shown that no particular category of candidate is

[11] Copies of the Wood report are available on request from the Bar Standards Board

disadvantaged, leading to inappropriate exclusion from the course. The threshold will be set very low. This is not a test used to select the best out of a large pool but rather to ensure that those unlikely to pass are not admitted to the many places available.

Similarly, it has been demonstrated that introducing the Aptitude test will not adversely affect diversity at the Bar. In fact, on the second pilot, there were high and low scorers within each identified ethnic group, and many ethnic minority students performed better on the Test than on the course. Instances of candidates performing poorly on the Aptitude test but doing well on the course were rare and included all categories. Evidence (in the form of time spent on the exercises) suggests that these candidates did not take the pilot seriously and did not apply themselves to the exercise. The percentage of applicants unable to meet the requirement will be very low and those individuals affected can take the test again once they have improved their skills.

As discussed above, raising entry requirements for the Bar Professional Training Course will doubtless raise the standards on the course as well as the students graduating from the Course. This is likely to increase the strength and effectiveness of the legal profession.

5.1.7 Increasing public understanding of the citizen’s legal rights and duties

The Bar Standards Board considers that requiring all BPTC applicants to take the Aptitude Test would have no effect, adverse or otherwise, on the public understanding of this regulatory objective. It would rather increase public understanding of entry requirements for training and provide additional reassurance.

5.1.8 Promoting and maintaining adherence to the professional principles, including maintaining proper standards of work; and acting in the best interests of clients

Promoting excellence and quality within the profession is a vital role of the BSB, as is ensuring that those who qualify as barristers have the right level of skills and knowledge to provide services to the public. This concept underlies the BSB’s role in regulating education and training for the Bar and ensuring a good quality educational experience which will inevitably lead to well qualified Barristers entering the profession. The requisite aptitude and skills are fundamental in ensuring the proper standards of work of those who are called to the Bar of England and Wales.

5.2 The Better Regulation Principles

5.2.1 Transparency

The Bar Standards Board has worked hard to be transparent through the entire review and consultation. A great deal of consultation and discussion has taken place to date, centred on the major review of the Course, before the conclusion that a universal testing system (for language, reasoning and aptitude skills) was the best way forward. Barristers, academic staff and the students themselves were all consulted and there was wide support for a stricter entry requirement to raise standards on the Course.

Consultation with other Approved Regulators on the requirement for all candidates to reach a specific minimum threshold on the Aptitude test will continue during the present consultation process.

5.2.2 Accountability

As the regulator for the Bar, the Bar Standards Board is accountable for the changes it implements. The Board has a wide range of plans to ensure that it remains accountable for the decision that it has taken to introduce a mandatory Aptitude Test. These include:

a. A programme of annual monitoring visits to all BPTC Providers where consideration will be given to effects the changes have had when talking to students, staff and management. Issues identified will be addressed as necessary.

b. The BSB employs External Examiners to assist with Quality Assurance by visiting individual Providers and attending examination boards. All External Examiners will be briefed to ensure that they particularly address the issue of any changes which have occurred due to the raising of the entry requirements, and ensure that they include any observations in their report for consideration by the Bar Standards Board.

c. The Bar Standards Board conducted a Student Perception of Course Survey in 2008, 2010 and 2011. There is currently a question on the survey seeking student views on whether they felt their experience was affected by the presence of weak students on the Course. This is monitored. The BSB will ensure that results will be analysed annually to monitor whether performance and student satisfaction improves on the course.

d. The Bar Standards Board will, by way of this report's conclusions, undertake various continual monitoring operations to ensure that the chosen test continues to be fit for purpose and to assess any effect it may have on the cohort.

5.2.3 Proportionality

As demonstrated above, the BSB carried out a thorough and careful review, with extensive consultation before determining the need for an Aptitude Entry Test for the BPTC. The conclusion reached was that some form of universal entry test was the only suitable way to ensure a fair and proportionate approach to the selection of suitable candidates. It is vital that minimum standards for entry to the course are set and monitored; particularly given the evidence over the years that self-regulation by Providers, according to the *minimum* entry requirements set by the BSB, has not been sufficient to resolve the problem. Notwithstanding the expense of the course, the evidence shows that students with insufficient aptitude and skills still seek to do the Course.

5.2.4 Consistency

The BSB believes that a requirement for all applicants for the BPTC to demonstrate a minimum level of aptitude for the course is by far the best method of ensuring that there is fairness in the selection and admissions process for the Course and that a consistent approach is adopted towards all students. This will also help to remove the possibility of inconsistencies according to Degree institution including those from overseas.

5.2.5 Targeting

The BSB considers that this is the appropriate stage of training at which to target the requirement for many reasons. It must take place prior to entry. The course is so expensive that it would be inappropriate for students to undertake the test after entry or to rely on outputs only. Although the BSB fervently believes in enabling access to training to a wide spectrum of applicants, it is nonsensical to suggest that all should be allowed to 'have a go', not least because of the effect the presence of weak students, particularly during small group discussions and interactive advocacy exercises, is likely to have on the learning experience and performance of the brightest and best who will progress to the Bar of England and Wales.

6 Evaluation

6.1 Constraints

There are constraints with these pilots and this report which must be addressed. Most of them have been mentioned earlier in this report; however it is beneficial to bring them together in one place. Many of these can only be addressed by the undertaking to continually monitor the effects of the Test after implementation.

We were unable to collect socio economic data for this pilot programme, due to the difficulty of collecting this data in any meaningful way for comparison with Test and Examination scores. While it is unlikely that the test will have any effect, adverse or otherwise, on socio economic factors of enrolment, the BSB undertakes to continue to explore ways to collect this data and to compare it.

It was especially interesting and enlightening to look at the results across different demographic groups. It is of course essential to do this as the test must not adversely affect any group unjustifiably or disproportionately. The constraint involved with collecting this data is that it was not compulsory for candidates to answer the questions relating to monitoring data and so the full data set was not available upon which to draw conclusions. While samples for gender and age groups were more than adequate, samples for some ethnic groups, disabled students and those without English as a first language were not large enough to be able to draw undeniable conclusions from these data sets. That said, meaningful conclusions were drawn from the data available and continual monitoring will be made much easier by the fact that all this information will be available from almost all students applying for the course.

Poor response rates to non-compulsory questions included those relating to Degree class and A Levels resulted in smaller samples for some analysis, and unreliable responses with respect to A Levels resulted in difficulty with some analysis. These questions are important to check whether likely course outcomes can be equally or better predicted by existing selection criteria. Although this data was not available for all students, it was clear that the Test scores were considerably better predictors than the existing criteria. In any case, this is something that the BSB will continue to monitor closely.

The Report Stage 3 identified that there is no grade for 230 students who took the Test originally but it is not possible currently to identify which are students who have since dropped out of the course and which are ones for whom scores have not been matched due to administrative issues. It is likely that this group contains a proportion of people who struggled with meeting the course requirements who might have been identified by the test.

It would have been beneficial to identify any links that may exist between English language test scores and BCAT scores. Unfortunately, although the question was asked when students undertook the pilot BCAT, no IELTS data was available. In the future, once the test is implemented, it will be beneficial to explore if there are links to discuss ways in which they can or should be amalgamated in the future. It is likely that comparisons would show that both tests are necessary as they test different skills and of course the BCAT does not have a speaking, listening or indeed writing element to it. Any comparison would only be a very crude measure, however, as it would have to be borne in mind that these comparisons would only be available for those who take the English language tests which is most likely to be predominantly international students (although the rule does not target these students). The BSB undertakes to continue to monitor and compare these scores when they are more readily available from a full pool of applicants and report appropriately.

As was briefly mentioned in the paragraph above, there is certainly merit to the idea of amalgamating elements of both the BCAT and the English language tests to create a bespoke language and ability test for use with all BPTC applicants. It is likely that such a test would be met with unanimous support as ensuring that only suitable students in terms of language and cognitive skills undertake the course. That said, such a test would need careful development and in order to ensure that the correct emphasis is placed on such a test, it would be important to run both tests concurrently for some time and conduct suitable analysis on performance.

6.2 Operational issues

Some details are yet to be decided by the BSB in relation to the active testing processes to be undertaken by Pearson. Discussions between the BSB and Pearson Vue began in 2010 looking further into details with the operation of the test. The following has been confirmed, but can be subject to change if investigation or consultation shows them to be unfavourable:

Registration

- The test will be called the Bar Course Aptitude Test, which shortens to BCAT.
- There will be no restrictions in relation to who is eligible to book and take the test. This means that people will be able to take the BCAT at any stage of their education or career, including after applying for the BPTC. The requirement will be that applicants must have scored the threshold pass before enrolling on the course, similar to the current English language rule and other entry requirements. Aptitude as measured by the test is resistant to change and therefore it is of no great matter when the test is taken.
- Tests will be delivered on demand, available at any time during the year (as long as testing centres are open).
- Registration will be available online. Candidates will be able to register and book a test at a location and on a date convenient to them.
- Estimated volume of tests per year is assumed to be a slightly higher number than applications for the BPTC as there will inevitably be some re sits undertaken as well as some students who are not yet applying for the BPTC taking the test at an earlier stage. To this end it is expected that there will be between 3000 and 4000 tests taken per year.
- The BSB will require Pearson to collect certain monitoring and personal data in order that the test can be effectively reported on in future and so that any records created on individuals can be merged with records which may be later held by the Bar Council.
- As recommended by the Independent Consultant the BSB engaged, the BSB will request that Pearson Vue provide candidates with an opportunity to undergo a

practice test to familiarise themselves with the styles of questions they may come across during the test.

Booking and taking the Test

- Locations in which the test is available will not be limited by the BSB; the test will be available to be taken in any centre in which Pearson Vue is able to deliver it. Pearson Vue has recommended that the same locations as the LNAT (Legal National Aptitude Test) uses would be appropriate, which the BSB has agreed with. This carries no additional cost implications for the implementation of the Test. The BSB does not at this stage seek to request Pearson Vue to expand their testing coverage as it is very wide reaching as it stands.
- Unfortunately there is a slightly higher cost of taking the test for international students due to higher cost to Pearson Vue of testing and processing results overseas. The BSB has explored the chance of a reduction with Pearson Vue but it has been confirmed that this cost is non-movable. As the cost accurately reflects additional costs which must be met by Pearson the BSB has confirmed that no unjustifiable discrimination will occur here.
- The BSB has reserved the right to view the Pearson Vue rescheduling and cancellation policy in full but generally agrees that 48 hours for rescheduling and 24 hours for cancellation (otherwise full payment taken) is reasonable.
- Candidates will be required to show valid photograph Identification in order to undertake a test in order to prevent fraud, further and detailed information will be included in their confirmation email when they book their test. Any candidate who does not have their Identification with them when they arrive at the testing centre will be considered a late cancellation and must rebook and pay for their test again.
- The test is currently operating as 40 questions; however the BSB’s Independent Consultant suggests that 50 questions would be most appropriate. One hour will be allowed to do the test (subject to any extra time allotted).
- Pearson Vue will ask the test takers to undergo a student satisfaction survey to which the BSB has requested access on an annual basis in order to monitor the operation of the tests.
- The test will only be available in English generally, although if a request is made for the test in Welsh, Pearson Vue has confirmed that these arrangements could be made. Such a request would require sufficient notice. No additional time will be available for candidates whose first language is not English or for those who are taking the test in any other country.
- Tests will include a non disclosure agreement to prevent fraud. Pearson Vue have a standard agreement, although the BSB has reserved the right to amend this to reflect that anyone found to disclose restricted information will give rise to serious questions about their fitness for call and may be expelled from the Inn and their Provider.

Re sits

- Although this report has been written based on the idea that an unlimited number of re sits will be available, it is worth exploring the pros and cons of this approach.
- As has been acknowledged earlier in this document, no selection tool is perfect. There will always be people who do not meet selection criteria who could have gone

on to be successful; this is also true of the current restriction on those who achieve a third class degree at university, for example. For this reason it is thought that the most fair option is to allow an unlimited number of re sits, to ensure that those who genuinely do have the aptitude for the BPTC are not prevented from doing so by, for example, a series of unfortunate events or repeated poor test taking technique.

- Further, it has been explained to the BSB that with tests such as the BCAT, practice and coaching can have a small effect, but after a certain time the effect will plateau. The introduction of the BCAT is naturally aimed at ensuring standards and therefore there would be no merit in allowing unlimited re sits if applicants could 'practice their way to a pass'. It is the view of the BSB that it is more important to ensure fairness by allowing an unlimited number of re sits, as the risk of applicants being coached sufficiently to achieve a pass is limited.
- Due to the fact that some candidates may leave taking the test until shortly before the commencement of the course, the decision has been taken to allow candidates to re sit the test as soon as they are able to book another sitting if they wish.
- One way of limiting re sits without placing a specific cap on the number of times the test can be taken is to require a period to elapse between re sits. For example candidates might be required to wait 2 months before booking another test session. This approach would help limit security risks with repeated retesting.
- In order to prevent fraudulent repeated test re sitting, the BSB and Pearson Vue will monitor re sit frequencies and will reserve the right to prevent a candidate from undertaking further re sits. This would obviously be investigated as necessary.
- It has been decided that there will be no re sit available on Tests that have been passed.

Adjustments

- Pearson Vue has confirmed that their testing centres are in a position to accommodate any reasonable adjustments as may be necessary. Industry norms include:
  - Extra time (25%, 50% and 100%)
  - Separate room
  - Reader
  - A combination of the above

Requests for reasonable adjustments for the test must be notified to the BSB by phone a certain amount of time before the test is set to be taken.

Since the test is considered a 'Power Test', it has a generous time allowance to allow all candidates to complete it without time pressure. To that end, there is already effective additional time built in. It has been suggested that therefore there would be no need to obtain evidence of disability to support requests for extra time as candidates who were not disabled who made dishonest requests for extra time would not gain an unfair advantage from an extended testing time. The BSB would reserve the right to check such evidence at a later date.

Requests for other, less frequent reasonable adjustments would require evidence and be dealt with on a case by case basis, to begin with. Pearson Vue would deal with physical disability adjustments.

Marks

- Marks will be given to candidates by use of a certificate on the day of their test. Their score will also be available online to the candidate and the BSB. Providers will be able to access a database showing which students have passed the test.

Security

- There is currently a bank of some 370 questions available for use. Several hundred new questions are being trialled to expand the question bank further. Pearson is currently using 335 questions from the bank to randomly generate 40 question tests for operational use. This will minimise the chances of any overlap of questions from one test to the next. It is however recommended that the test for the BPTC be used with 50 questions.

Question Bank Replenishment

- New questions must continue to be developed so that they can be added to the bank and overused questions can be retired. On an ongoing basis it may be desirable to require candidates to complete a slightly longer test (in terms where some of the questions are new trial questions under development). A 60 item test would allow each candidate to complete 10 developmental questions. Even with the addition of these extra questions, the test should not take longer than 40 minutes to complete for most candidates, although one hour will be allowed.

7 Conclusion

7.1 Summarising the main findings of the pilot

1. The correlation between test scores and course examination results was 0.51 across the whole sample. This is a very strong result and highly statistically significant. A much smaller correlation would have supported the use of the test in selection.

2. When looking at the correlation with examination results compared with that for Degree grades or A Level results, the BCAT is the best single predictor of course outcomes and in fact it can predict course outcomes better than Degree grades and A Levels.

3. Although it might seem at first look that discrimination may occur as certain groups, particularly some ethnic groups, perform less well on the BCAT, this is not in fact the case as these differences are reflected in course outcomes to the point that the test actually over predicts performance in the course for the groups that score least well on the BCAT.

4. A cut score of between -1 and -1.34 is recommended in order to improve failure rates on the course, improve the educational experience of those on the course and ensure that the majority of people who are likely to fail the course are identified at an early stage and do not spend huge amounts of money on a course they cannot pass. This range of cut scores is also low enough to minimise the ‘false negative’ effect of some students who may do well on the course failing the test. These students would of course have an unlimited number of re sits. Where there are score differences between demographic groups, this range of cut scores ensures selection ratios do not breach the four fifths rule.

5. The test can reduce the number of failures and referrals on the course, it can also improve the proportion of students who achieve the highest grade and pass the course at first sitting.

7.2 The Desired outcome

The desired outcome from the proposal to implement the BCAT for all prospective BPTC students would be the improvement of standards on entry, and therefore exit of the BPTC. By using a test universally we will be able systematically to identify students who are likely to fail the course and they will be prevented from undertaking it, thus saving them wasting their own time and money.

The proposed alteration is expected to increase student satisfaction on the BPTC as students with low aptitude will not be on the course to risk affecting other students’ experience. Not only is it expected to increase satisfaction of students but also tutors and course leaders, thus improving the student experience all round. This will be monitored by meeting with students during monitoring visits, encouraging our External Examiners to do the same and comparing the results of the Student Perception of Course Survey from previous years with those in subsequent years, after the rule change.

It is expected that overall the failure rate on the BPTC will fall. This will be due to the fact that there will be a lower proportion of students on the course with a propensity to fail and also that classes will not be slowed down by students with a low aptitude. This will be monitored by comparing the scores and statistics during the years after the rule change with those from previous years. The new BPTC is (as with the previous BVC) carefully monitored annually by means of scrutiny of the Annual Monitoring reports, by Providers, by annual visits to each provider/site by the BSB and by obtaining feedback from students through discussions and questionnaires/surveys. A review, consisting of the cumulative feedback and findings of annual monitoring is intended to be carried out after 3 years of the new course and requirements. The effects of the BCAT in raising the standards on the course, improving the learning experience of successful students, and increasing the quality of graduates is anticipated and would be shown by results and feedback from various stakeholders.

It is expected that the implementation of this proposed rule change will also assist the Legal Services Board in achieving the Regulatory Objectives due to the reasons outlined elsewhere in this report.

8 Consultation Questions

1	**Do you consider that current entry requirements on the BPTC need to be changed?**
2	**Do you agree with the rationale for implementing an additional entry requirement for the BPTC in the form of a universal Aptitude Test (BCAT)?**
3	**Do you consider that the introduction of the BCAT is justified by the data presented in this report?**
4	**After looking at the results of the pilot tests, do you consider that the BCAT will reliably identify students who are likely to fail the BPTC?**
5	**From looking at the evidence in this report, what cut score would you consider to be most appropriate?**
6	**Do you agree that an unlimited number of re sits for the BCAT should be available (subject to anti-fraud frequency monitoring)? If not, how many do you consider to be appropriate?**
7	**After looking at the results of the pilot tests, do you consider that the introduction of the proposed BCAT would have a disproportionate effect (either positive or negative) on any particular group compared with others?**
8	**Are there any negative impacts that have not been identified in the equality impact assessment?**
9	**Do you consider that entry standards on the BPTC could be made more rigorous in a way other than what is suggested in this paper? If yes, please expand.**
10	**Please insert any other comments on this consultation document here.**

Comments received during BVC Review & Consultation 2007-2008

Provider comments

- An aptitude test or interview should occur before starting the course (Nottingham Law School)
- The design and administration of such a test would be the key to its success and may prove to be a less blunt instrument than relying on purely academic selection criteria. Consideration would have to be given to the administration and cost of any such test (Cardiff University)
- The School answered that it was difficult to set such a test, and that it would be a costly to run the test and to deal with the academic appeals that would doubtless arise. ICSL

The only negative response to the proposal to consider an aptitude test was from Manchester Metropolitan University:

- Again we consider the proposal to require students to take an aptitude test as an unnecessary measure to adopt. Students already sit such a test – it is called a Degree! (MMU)

Comments by Students

- Students were in favour of an aptitude test for the course and felt that a charge for this of £100-£200 for the test would not put people off, as it makes little difference to the overall price of the course. (College of Law Birmingham students)
- The students felt that the entry standards for the course need to be raised and suggested a more comprehensive application form. They state they would have been happy to complete an aptitude test, depending on the costs involved. (BPP students)
- The question of entry standards had also been discussed with students who recognised that there were some significantly weaker students on the course. The course was more about skills than academic ability so performance was not (in the view of the students) related to Degree classification. (Students BPP Leeds)
- English language problems and diction, accents etc should be addressed (including native speakers) due to impact on others' learning experience. A 2:1 entry requirement was supported, with additional criteria for some waivers for those with 2:2s. An aptitude test might be worth considering. (Lincoln's Inn student activities committee)

Consumer, Diversity Groups and Law Society

- it is important to raise the standards of students not only to control numbers but to reduce the number of weak students and attract better ones; and to reduce wastage. the threshold pass for students on the course could also be raised. (Consumer Panel)

- There was support for the concept of a universal aptitude test, although there are logistical difficulties and careful monitoring will be needed (Equality & Diversity Committee)
- Strong support was expressed on behalf of the Law Society for the proposal. A similar approach was being considered by the Law Society Education & Training Committee at the time. (Dec 2009)

Practitioner Committees and Specialist Associations

- An aptitude test could be reinstated to select appropriate candidates (Employed Bar Committee)
- MCQ tests might be used as elements of a pre BVC aptitude test (Young Bar Committee)
- It was agreed that there could be competition issues if numbers are restricted, but that entry requirements (for example Degree classification, language and/or other aptitude testing) should be introduced to ensure standards.... entry requirements should be rigorously used to ensure standards; additional testing systems could be used (BACFI)
- There was more support for the idea of some sort of aptitude test, though only if it were conducted centrally, probably by the Inns, rather than by the course providers themselves. (Family Law Bar Association)
- BVC should be pitched at a higher level so it would be more respected. ...strong message about the need for high level language requirements (Tech Bar)
- All would-be applicants have completed at least 4 years law studies, and many have at least one doctorate. After the 4 years+, applicants have to pass the CRFPA (Centre Regional de Formation a la Profession d'Avocat) entrance examination. This nationally regulated examination involves 2 five hour examinations, a three hour examination and a dossier: no simple aptitude test! (Ecole de Formation Professionelle des Barreaux du resort de la Cour d'Appel de Paris)

Inns of Court

- concerns about the intellectual and linguistic skills of students, hence support for universal 'LNAT-type' aptitude test (Gray's Inn education committee)
- an aptitude test would be helpful but would be expensive and difficult to administer (Middle Temple Education Committee)
- The general consensus was strongly in favour of an aptitude test whilst recognising that careful thought would be required to devising an aptitude test which addressed the requirements for practice at the bar and diversity issues.. There would be an inevitable time-lag before it could be introduced. It was proposed that a 2:1 entry could be used as an interim measure. It was agreed that an aptitude test designed to assess whether a student had the necessary skills to practise at the Bar and a high level BVC would promote excellence in the profession. The general consensus is strongly in favour of an aptitude test. (inner Temple education Committee)
- The possibility of applying an aptitude test should be explored COIC (March 2008)
- MCQ tests might be used as elements of a pre BVC aptitude test (Young Bar Committee)

- there was a feeling that this might be difficult to administer, and that it would be difficult to devise such a test and monitor its appropriateness and effectiveness as a selection too (Lincoln' Inn Education Committee)

BSB discussions

- Aptitude tests, it was agreed, are likely to be tricky to implement, although they might have advantages in forming A Level playing field between those of differing school and university backgrounds. The use of such tests in the USA for law schools indicated that it might be possible. (minute of BSB discussion, March 2008)

Initial analysis of Watson Glaser Critical Thinking Test: First Pilot Results

17th August 2010

Initial analysis of Watson Glaser Critical Thinking Test (WGCTT) Pilot Results

During the first half of 2010 students on the BVTC were encouraged to volunteer to take the WGCRT. 182 Students completed the test. These are self selected sample and therefore may not be representative of the student body as a whole. In particular it may be that weaker students were more reluctant to take part.

A summary of the demographic make up of the sample:

- 43% Female
- 55% White
- 84% English Primary Language
- 65% Aged under 25

Table 1 provides an overview of the results of the pilot with the test. It can be seen that all the indicators are within a desirable range.

Table 1:

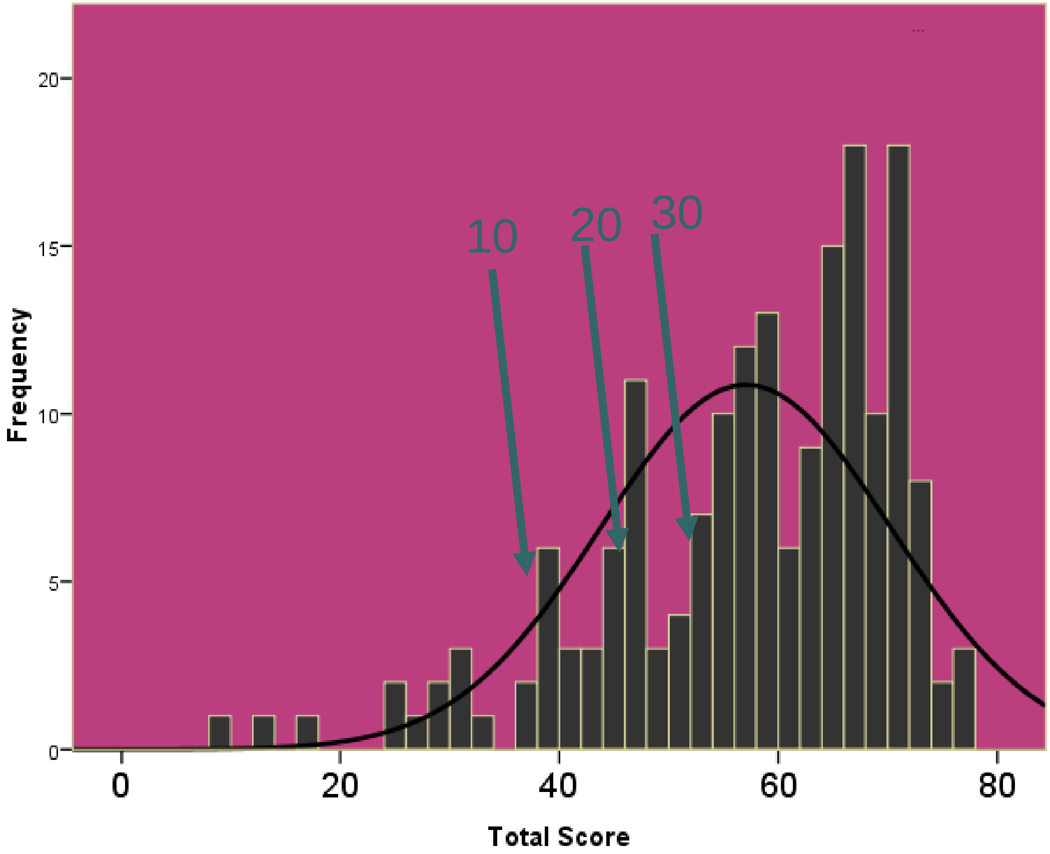
Indicator	Finding	Desirable
Mean	57 (out of 80)	40-65
Std Dev	13	8 < sd < 15
Reliability	0.93	> 0.80
SEM	3.6	< 5
Completion Time Average / Max	31/49 minutes	Max < 1 hour

Score Distribution

Figure 1 shows the distribution of scores on the WGCRT. It can be seen that there are a good range of scores, even in this volunteer sample of course students. There are some lower scores that would allow the use of the test to select out poor students. The arrows show where the cut score would need to be to reject 10%, 20% and 30% of the sample. However it is important to remember that lower scores in this sample may be due to poor motivation among students rather than poor ability. On the other hand it is likely that there will be more people of low ability among applicants than students on the course.

Overall this distribution shows desirable properties but should be verified on an applicant sample.

Figure 1



Test Completion and Timing

Table 2 provides the results regarding completion rates with respect to the 80 questions in the test. Again it should be remembered that motivation is always an issue with a volunteer sample and that low motivation can lead to both slower performance through lack of effort and also to a speedy slapdash approach. Overall most people were easily able to finish the test in the time allowed and most that did not finish only missed a few questions.

The average completion time was about half an hour which is easily acceptable; even the slowest student completed in less than an hour. Figure 2 shows the range of completion times. It is likely that those completing in less than 20 minutes were not trying to achieve a maximum score.

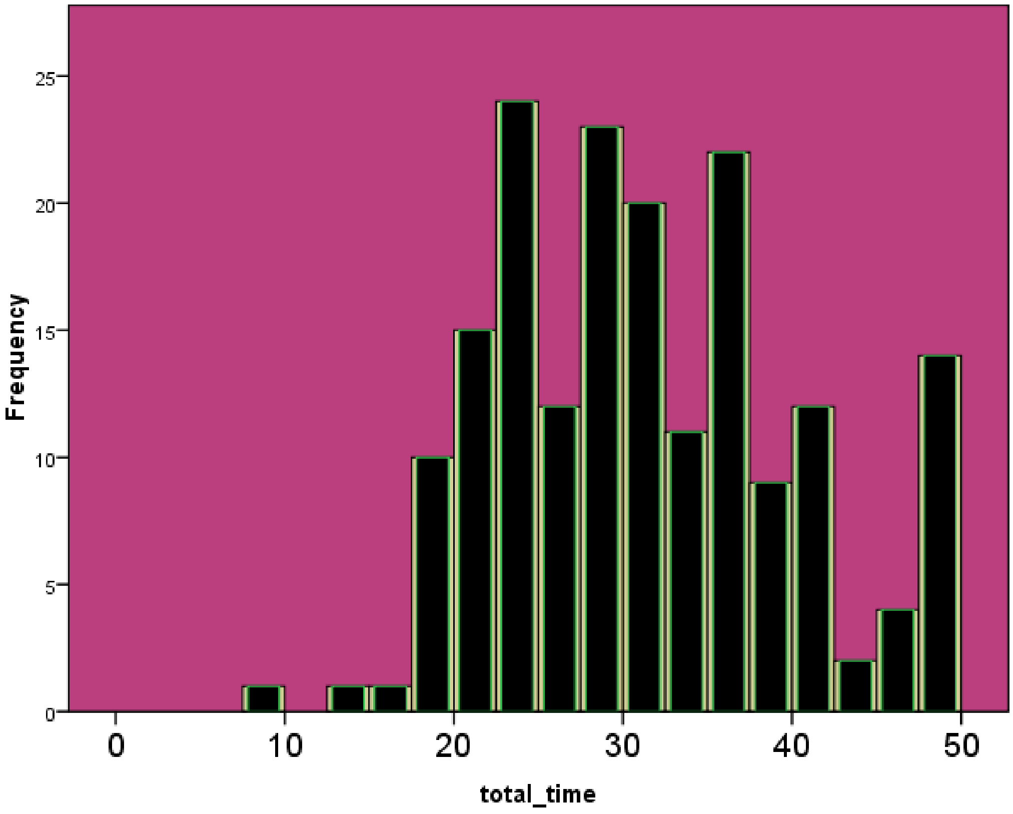
The completion is well within an hour and allows the use of a longer test in the next pilot without exceeding one hour of testing time. This will allow more questions to be pre-trialled.

Table 2

	Percent of sample
Complete all items	79%
Less than 5 missed (<5%)	10%
6-16 missed (6- 20%)	3%
17-32 missed (21-40%)	4%
33-55 missed (41-60%)	4%
> 55 missed	0.5%

Figure 2

Time to Complete (mins)

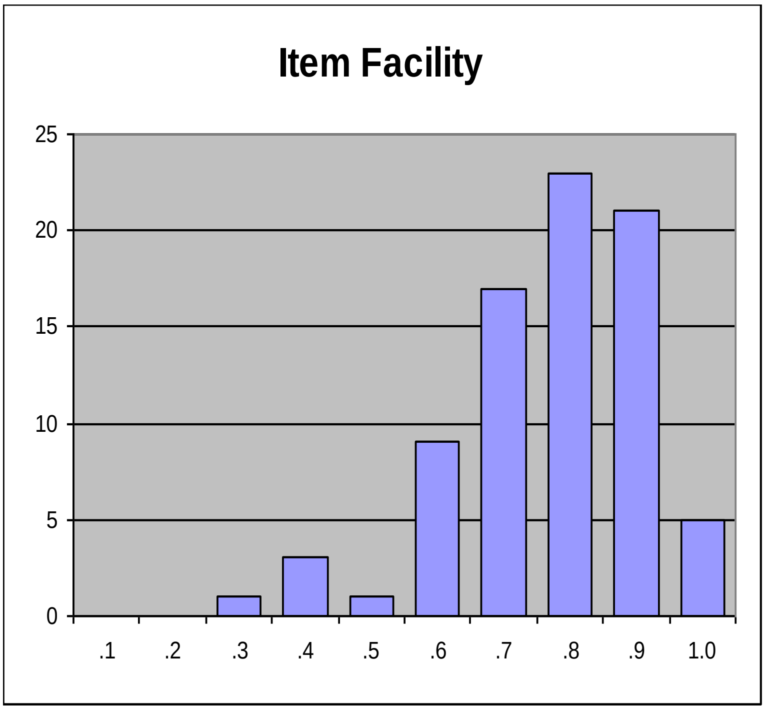


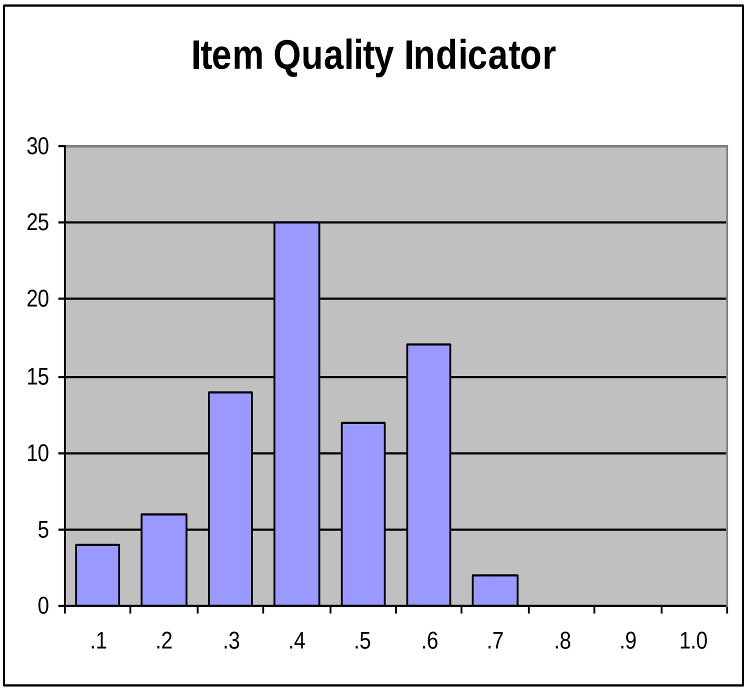
Item Indicators

It is important that not only the test overall is effective but that all of the questions (items) are of good quality. A range of difficulty of questions helps to differentiate between candidates. Figure 3 shows that item facility ranges from questions only 30% answer correctly up to those that nearly all students answer correctly. The majority have between 60% and 90% answering correctly. This is appropriate for a test where many questions have only 2 options and 50% correct is expected through guessing alone.

The second histogram in figure 3 shows the quality of questions in terms of their consistency with overall score on the test. Values above 0.2 are desirable although very easy questions do not always achieve this level. It can be seen that nearly all questions are in the desirable range with many very much above it. This shows that the questions are working well for the group. It will be necessary to review those questions with a lower index. They can be removed from the question pool if found to be unsatisfactory.

Figure 3: Question Indicators





Group Comparisons

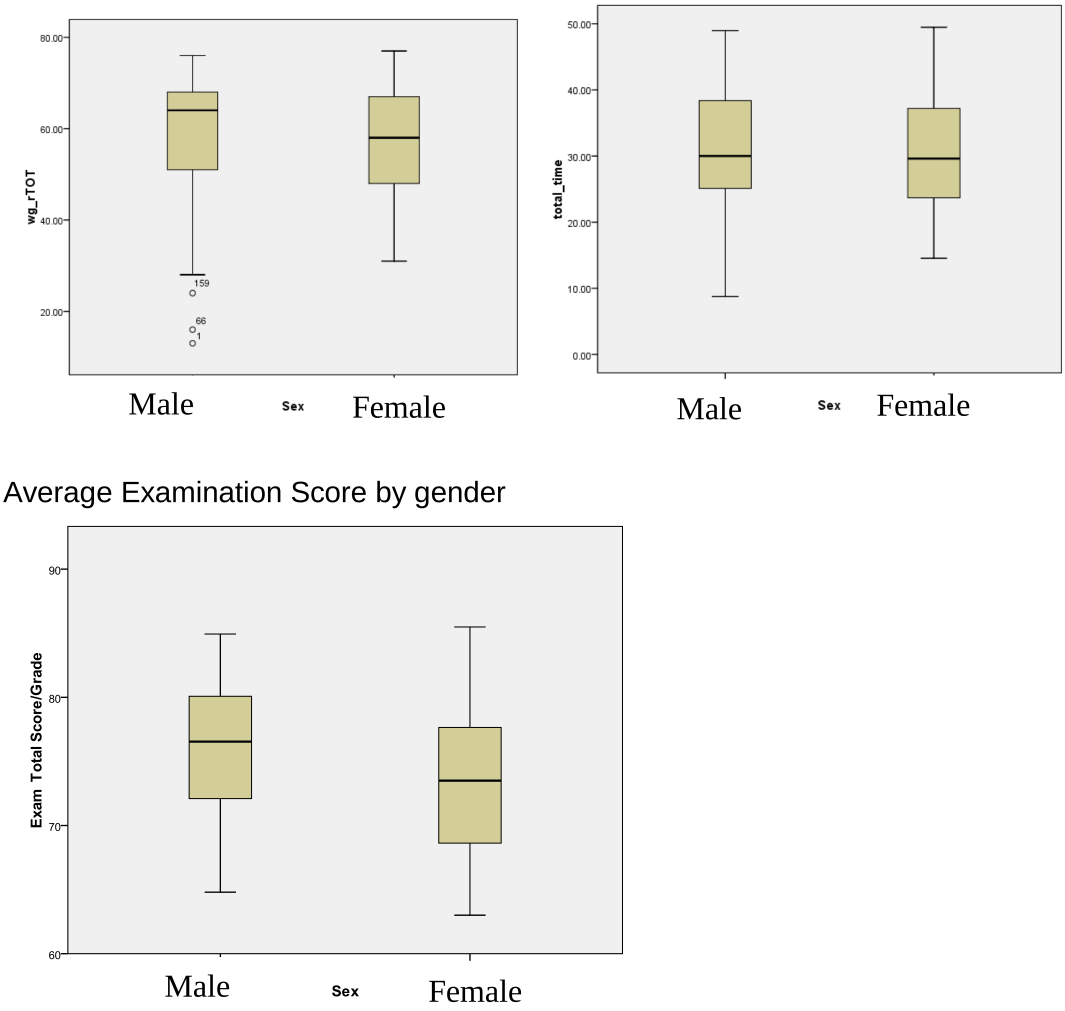
Figures 4 to 7 provide a comparison of score ranges for various background variables. Again it should be remembered these are self selected volunteer samples and this affects the ability to generalise to the full student population. There is greater difficulty generalising to the applicant group and results in use may be somewhat different. Continued monitoring and analysis is recommended.

The results are presented as box plots. The box in the middle of the column shows the middle 50% of scores. The line through the box is the median (middle score). The whiskers above and below show the range of the remainder of the scores although exceptional outliers are shown as separate circles.

When comparing 2 groups, if the two boxes are on about the same level, then the score ranges are similar. If the median lines are also on the same level then the groups are very similar. Differences in the whiskers are less important but indicate a larger or smaller range of scores for one group.

There is little difference in score or time taken by gender. However examination results for men are slightly but significantly greater than those for women.

Figure 4: Score, examination results and completion time by gender



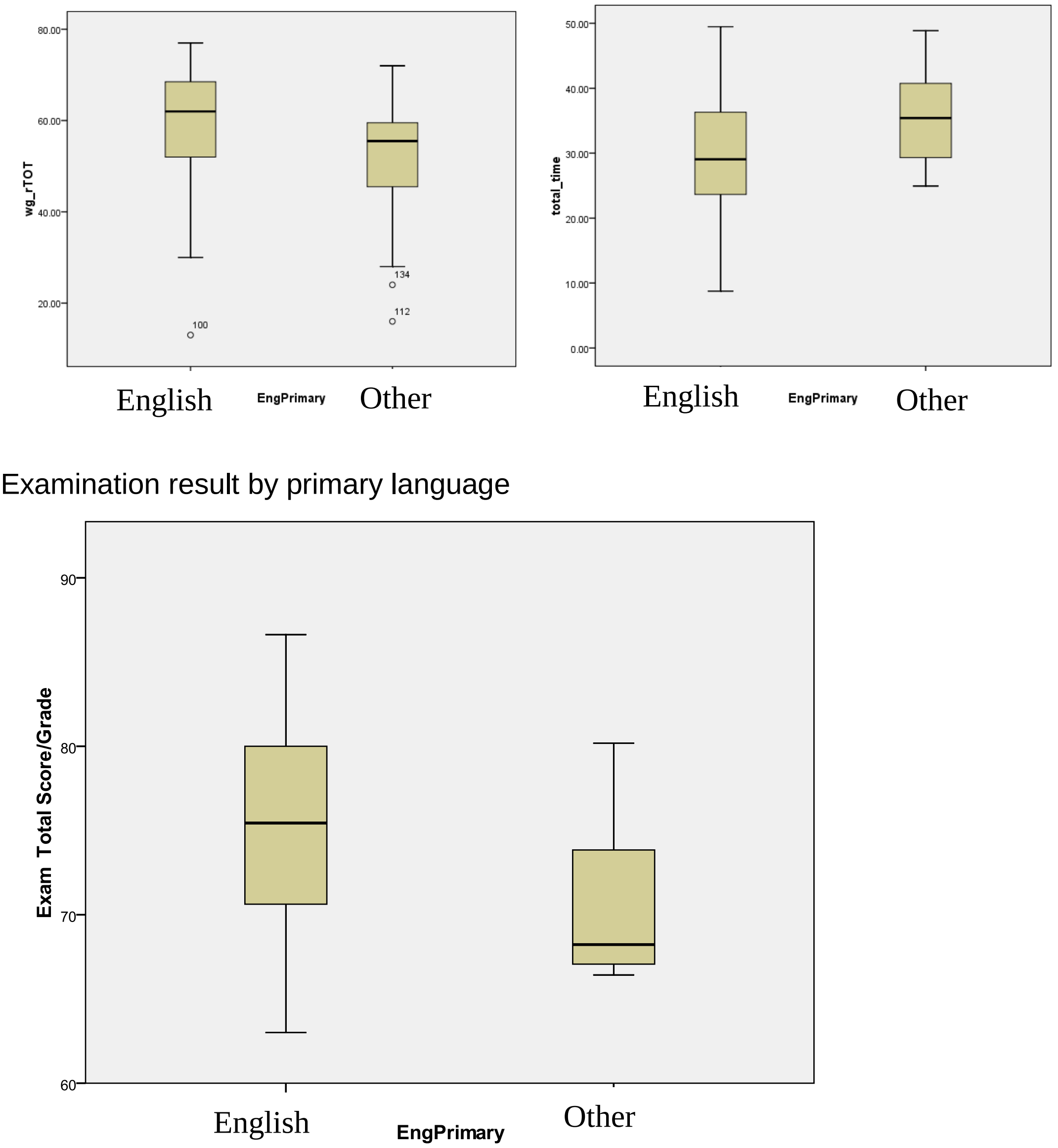
Score by gender

Time taken by gender

Average Examination Score by gender

Figure 5 shows that those with English as a primary language perform better and faster than those for whom English is a second or later language. The same difference is seen in examination performance which is also substantially poorer than for primary English speakers. All these results reach statistical significance. The difference in examination score is about one standard deviation and the difference in test scores is a little smaller.

Figure 5: Score, examination results and completion time by primary language

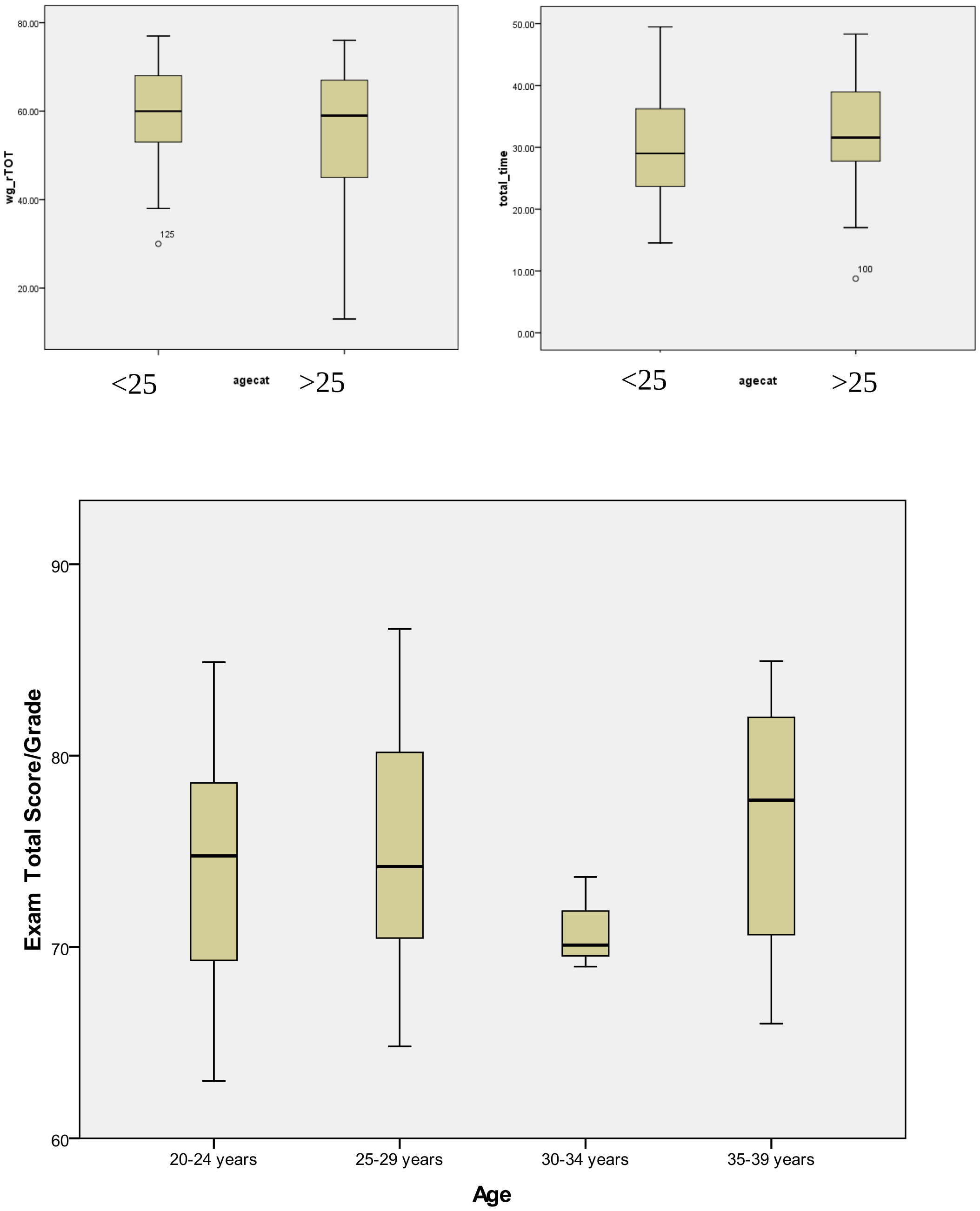


Score by Primary Language

Time taken by primary language

Examination result by primary language

Figure 6: Score, examination results and completion time by age
Score for under and over 25s Time taken for under and over 25s



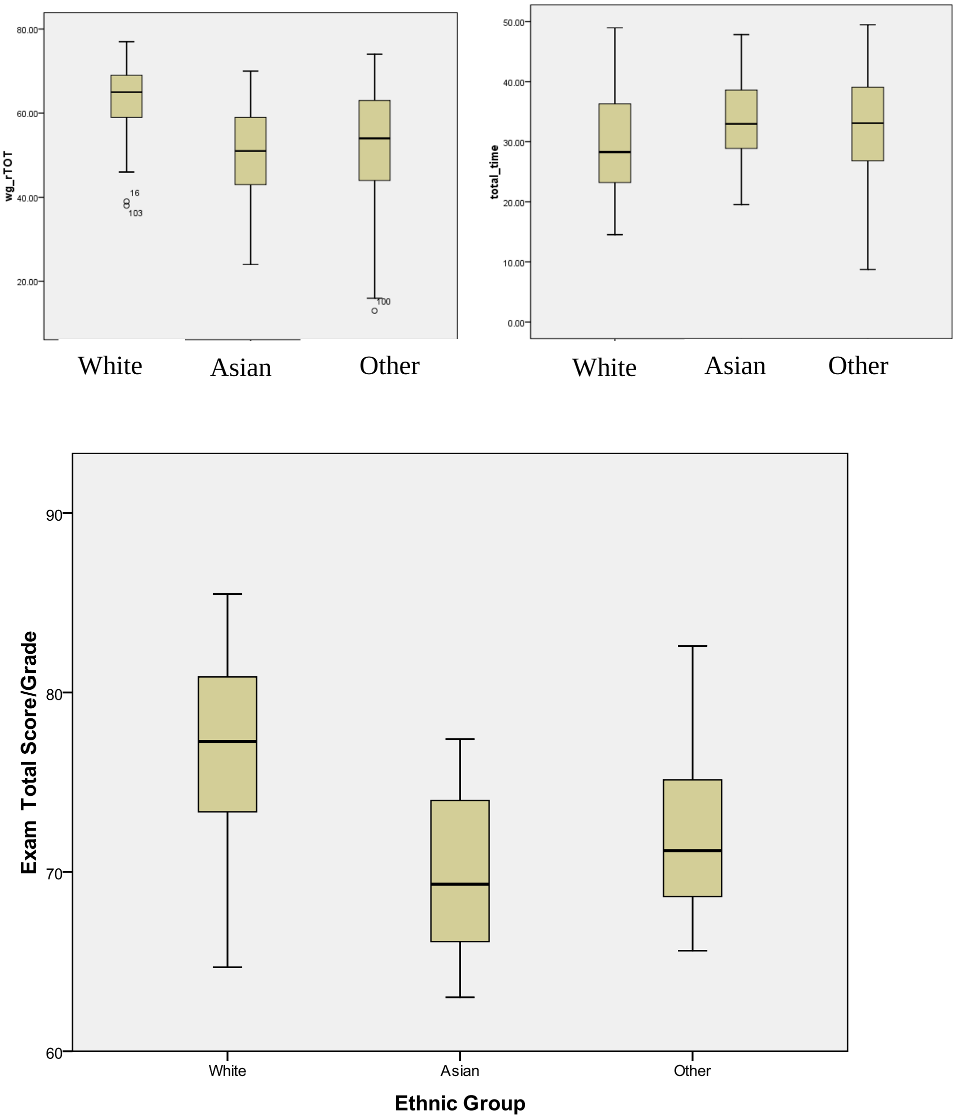
There is a small, but statistically significant difference in performance by age with younger students performing better on the test. There is more variation in older students. There is no significant different in examination results. This finding should be monitored in future samples. Age may well be confounded with other background variables. In particular older students may tend to have taken a different academic route to prequalification for the BVTC.

Figure 7 shows larger differences by ethnic group with over one standard deviation difference in both examination scores and test scores between White candidates and others.

White candidates have fewer very low scores on the test but this pattern is not seen in the examination results. White candidates also complete the test a little quicker. These results are somewhat similar to those for primary language and of course a greater proportion of the White students have English as a primary language. Thus there is a confounding of the two factors.

Figure 7. Score, examination results and completion time by Ethnic Group

Score by Ethnic Group Time taken by Ethnic Group



Validation: Relationship between Test Scores and Examination Results

In May/June 2010 most of the students in the sample sat their final examinations and results have been obtained for the majority of these students.

This note provides the initial results of the comparison between course results and scores on the WGCT. These results should be seen as provisional pending the finalised results of the whole sample.

WGCT scores range up to 80 with the minimum obtained in the trial being 9.

Course results are classified as Outstanding, Very Competent, Competent or Referred. At this stage some students may be waiting to re sit some papers in September for various reasons.

Table 3 shows the average WGCT score for candidates in each outcome category. It can be seen that there is a clear relationship between WGCRT scores and the result with the students achieving Outstanding scoring an average of 24 more points on the test (2 standard deviations) than the Referred group with the Very Competent and Competent have intermediate average scores.

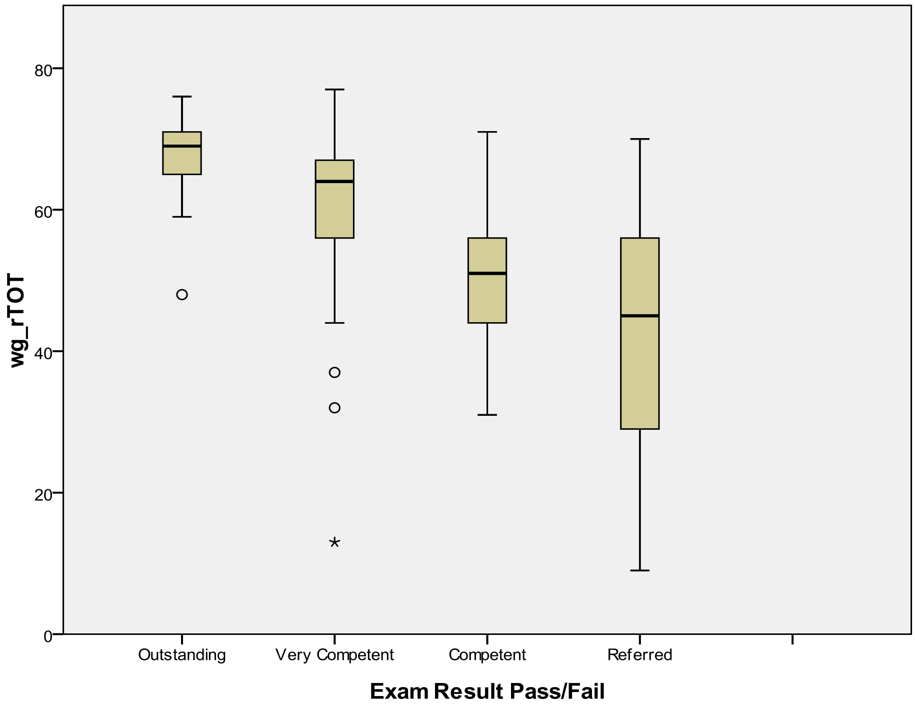
Table 3: Average WGCRT scores by Result

BVTC Result	N	Average WGCRT score	Standard Deviation WGCRT	Minimum WGCRT	Maximum WGCRT
Outstanding	18	67.6	6.6	48	76
Very Competent	83	61.2	10.1	13	77
Competent	31	50.7	10.1	31	71
Referred	24	43.2	16.5	9	70
Total	182	57.1	13.2	9	77

Figure 2 is a box plot showing these results. In this plot each column shows the WGCRT results for one category of BVTC outcome.
It is clear that there is a strong relationship between outcome and test scores for those that have final scores. The referred group has nearly all the lowest scoring individuals but also contains some higher scoring students. This may reflect the variety of reasons which lead to students needing to re sit examinations.

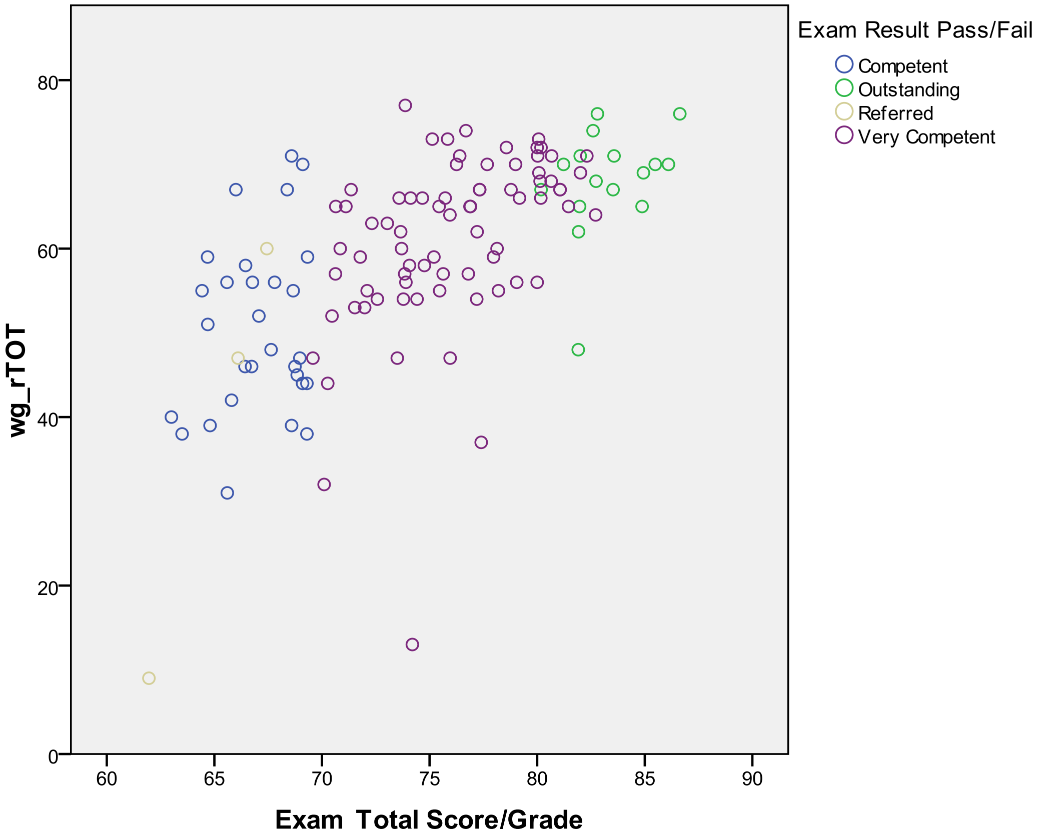
One student rated ‘Very Competent’ had a very low WGCRT score. This student completed the whole test in 13 minutes when the average time taken was over half an hour. This suggests the student may not have invested much effort in completing the test and the result may underestimate his/her ability.

Figure 8: Box plot of WGCRT scores and course outcomes



As well as the category of result the average examination score was available for 123 students. Average scores ranged from 62 to 87 with a mean of 74.6 and a standard deviation of 5.0. Figure 9 shows the scatter plot of scores on the WGCRT compared to the average examination score overall. The points are coloured according to the final Examination Result. Again the clear relationship between the two sets of scores is evident. However there were few scores available for candidates who were referred and therefore the plot does not show what score range, if any, is typical of those who are likely to fail the course.

Figure 9: Scatter plot of average examination results and WGCRT scores



The relationship between the test and examination scores can also be expressed as a correlation. This is a number between -1 and +1 which reflects the strength of a relationship; the higher the absolute value, the stronger the relationship. For selection purposes correlations of 0.3 and above are desirable but those above 0.2 are still useful. The correlation between the WGCRT and examination scores is 0.62. However there are 2 outliers in the data which are cases with very low test scores. Outliers can have a big impact on correlations so the result was recalculated without these cases. It remained high at 0.60. Both these results are highly statistically significant and indicate that the WGCRT will be very effective in identifying how well applicants are likely to perform on the course.

However it still remains the case that the current data set does not include many students who may fail the course. Only 24 of those who completed the test have been referred and some of those may do well at re sits. It is therefore not clear from these results specifically how effective the test will be in identifying students likely to fail or what level of scores on the WGCRT is associated with a higher probability of failure.

Table 4 shows how the result category relates to the demographic variables. It shows the number of people in each category from each group. The numbers are quite small in some categories so generalisations should be made with care; however it is clear that white students have better results than other groups. 82% of white students achieve Very Competent or Outstanding whereas only 40% of Asian and other students do. There is a similar contrast between those with English as a primary language who have a 72% success rates and those with English as a second language who have only a 20% success rate. This latter is quite a small group and a larger sample is needed to be sure of the success rate for this group.

There is little difference in results between men and women or older and younger students.

Table 4: Breakdown of results category by candidate background

	Outstanding	Very Competent	Competent	Referred	Total
White	13	43	9	3	68
Asian	0	11	13	8	32
Other	1	9	5	5	20
Total	14	63	27	16	120

Language	Outstanding	Very Competent	Competent	Referred	Total
English Primary	15	63	20	10	108
Other	0	3	6	6	15
Prefer not to say	0	1	2	0	3
Total	15	67	28	16	126

Sex	Outstanding	Very Competent	Competent	Referred	Total
Male	8	31	8	9	56
Female	5	37	20	7	69
Prefer not to say	1	1	1	0	3
Total	14	69	29	16	128

Age	Outstanding	Very Competent	Competent	Referred	Total
Under 25	5	48	17	11	81
Over 25	9	17	10	5	41
	14	65	27	16	122

Conclusions

These results show that the test is performing well with this group and that there is a strong relationship between the test and examination results. This bodes well for the future use of the test.

What this study cannot show is whether the test can identify applicants likely to fail because they are not included in the sample. Extrapolating from the current data suggests the test will be able to do so, but there is always a danger of extrapolating results outside the bounds of a sample.

Further analysis is needed when final results are available for the full sample.

The Evaluation of the Aptitude Test for the Bar Professional Training Course (Second pilot)

Report 1: Monitoring

19th August 2011

Background

The Bar Standards Board is currently piloting a critical reasoning test for use as part of the recruitment procedure for the Bar Professional Training Course. It is important that before the test is implemented all the appropriate research and checks are carried out to ensure that the test is effective, appropriate, fit for purpose and is fair to all candidates.

A small pilot was undertaken with the 2009/10 student cohort which suggested that the proposed test had the potential to be an effective predictor of course outcomes. Good prediction is necessary for the identification of candidates who do not have the capacity to complete the course successfully. The pilot also identified some differences in performance between groups.

In order to further evaluate the use of the test a larger pilot was undertaken with the 2010/11 cohort. This report provides the results of monitoring comparisons for this group. Further reports will deal with prediction of course outcomes and use of the test in practice.

Sample

All current students on the BPTC were asked to attend a Pearson Vue testing centre to take the test during November 2010. The students were asked to take the test relatively early in their studies to minimise the impact of the course on their performance.

1501 students attended the centres and completed a version of the test. Just under 1000 completed the demographic information requested.

A summary of the demographic make up of the sample:

- 55% Female
- 56% White
- 92% English Primary Language
- 66% Aged under 25
- 9% Have a disability

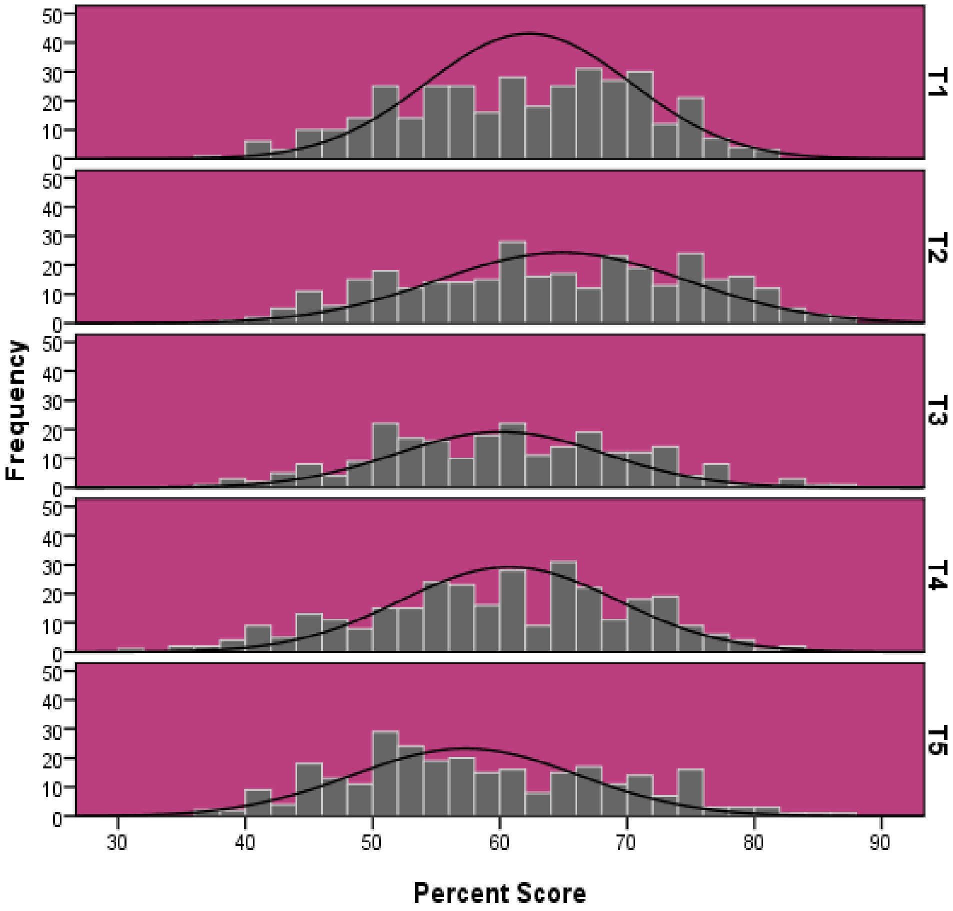
Score Distribution

For a test to be effective it should differentiate well between high and low scorers. This means that a range of scores are required. The final distribution of scores will depend on the length of the test chosen. For the current test students completed one of 5 versions of the test which were each longer than the test for operational use.

This was to enable the piloting of a large pool of questions which would support multiple test versions.

Figure 1 shows the distribution of scores on each of the 5 versions. Because there were differences in the number of questions, scores are expressed as percentage correct. It can be seen that there are a good range of scores. There are some very low scores that would allow the use of the test to select out poor students. In interpreting these results it is important to remember that lower scores in this sample may be due to poor motivation among students who completed the test rather than poor ability. For these students there was no important outcome depending on performance on the test so they may not have performed at their best. On the other hand it is likely that there will be more people of low ability among applicants than these students who have passed the selection criteria to gain their place on the course. Where colleges accept the majority of applicants, results can be expected to be similar but if colleges are more selective results for applicant groups may be somewhat different.

Figure 1: Score distribution for 5 pilot test versions



In order to understand the relative performance of the group their results were compared to a standard sample of graduate applicants to jobs who had taken a version of the test provided by the test publisher. Table 1 shows the results. The results use a metric with an average of zero and a standard deviation just below 1. This is developed using Item Response Theory and has the advantage that scores are comparable even when individuals have taken different test versions. The results show a similar standard of performance with the performance of the BPTC students just below that of the graduates. There is more variation in the BPTC students' scores. There is likely to be greater variation in an applicant sample.

Table 1: Comparison of BPTC students with other Graduates

	Bar Students	Graduate Comparison Sample
Mean	-0.17	-0.12
Standard Deviation	0.88	0.61

Group Comparisons

Table 2-9 and Figures 2 to 8 provide a comparison of score ranges for various background variables. It should be remembered the sample is made up of pre-selected students and may not entirely generalise to an applicant group. Continued monitoring and analysis is recommended.

The results are presented in a table and a figure for each comparison. The table shows the mean and standard deviation for each group and the raw and standardised differences between groups. Scores in the tables are expressed in the IRT metric. This has a mean score of around zero so is a good basis for comparing performance. The standardised difference is Cohen's d statistic. Because of the large samples results are likely to reach statistical significance at the 95% level even where differences are quite small. The d statistic provides an estimate of the size of the difference with values below 0.3 categorised as small, around 0.5 in the mid range and above 0.8 as high.

The figure is a box plot. The box in the middle of the column shows the middle 50% of scores. The line through the box is the median (middle score). The whiskers above and below show the range of the remainder of the scores, although exceptional outliers are shown as separate circles.

When comparing 2 groups, if the two boxes are on about the same level, then the score ranges are similar. If the median lines are also on the same level then the groups are very similar. Differences in the whiskers are less important but indicate a larger or smaller range of scores for one group.

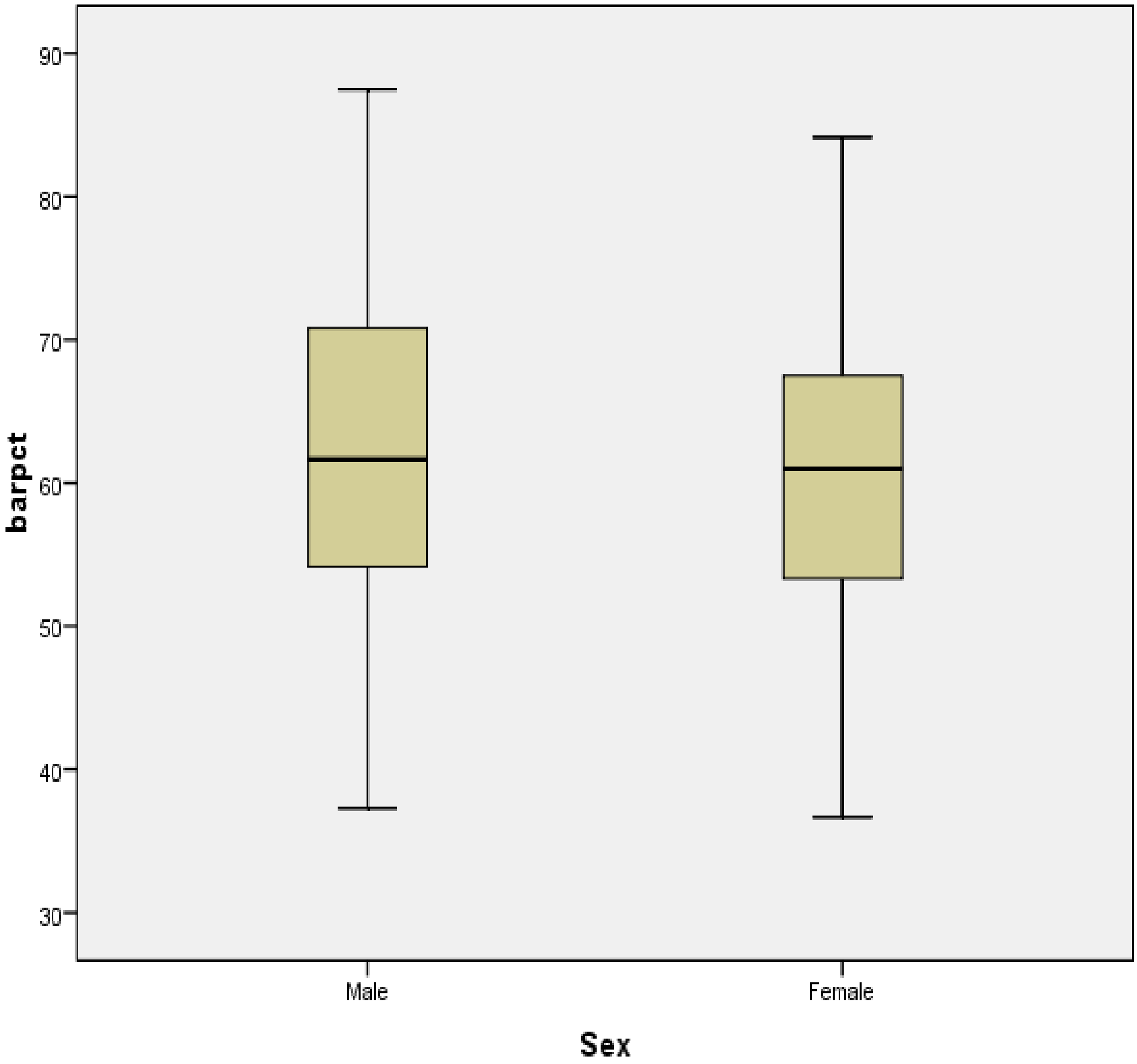
Scores by Gender

There is a small difference in scores with men scoring marginally higher than women but the difference is small although it does reach statistical significance.

Table 2: Comparison of Test Results by gender of student

	Mean	Standard Deviation	Sample
Male	-0.05	0.90	449
Female	-.20	0.82	545
Difference	.15 (raw)	0.18 (standardised)	994

Figure 2: Score by gender



Scores by Primary Language

Figure 3 shows that those with English as a primary language perform a little better than those for whom English is a second or later language. A similar difference has been seen in examination results in the past although these have not yet been analysed at the time of writing. These results reach statistical significance. The difference in test scores is a little larger than for gender but still in the small range.

Figure 3. Scores by primary language

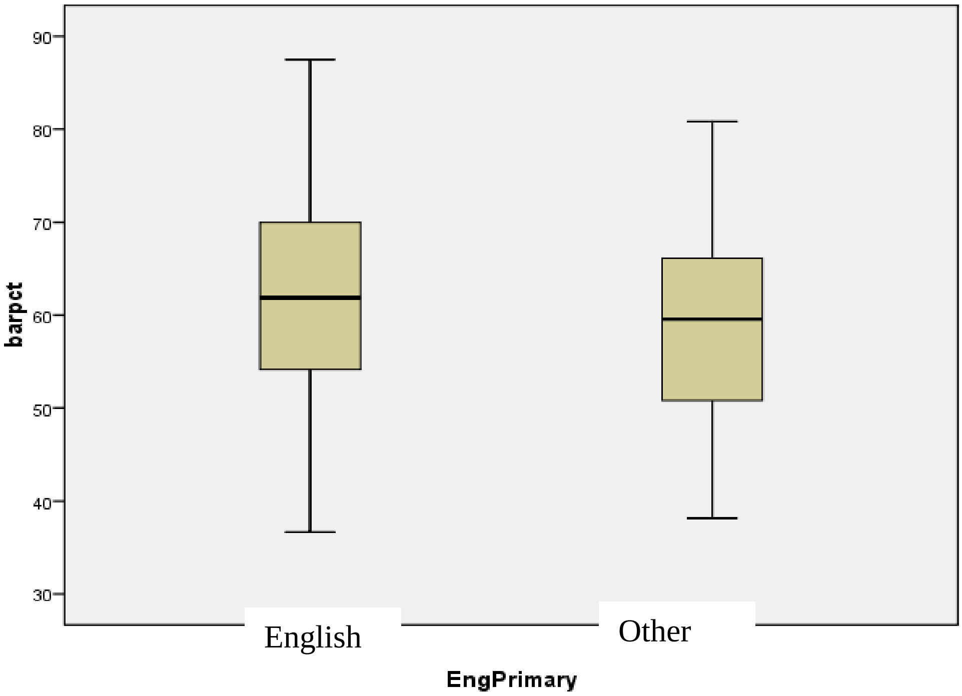


Table 3: Comparison of Test Results by primary language

	Mean	Standard Deviation	Sample
English	-0.09	0.86	867
Other	-.33	0.79	80
Difference	-.24 (raw)	0.28 (standardised)	

Scores by Age

There is some decrease in scores with age and the difference is most marked between those in their late 20s and those in their 40s but the differences overall do no reach statistical significance and may just be an artefact in the data. The sample size is quite small for the older groups. The largest d statistic is at the top of the small effect range.

Figure 4: Scores by age

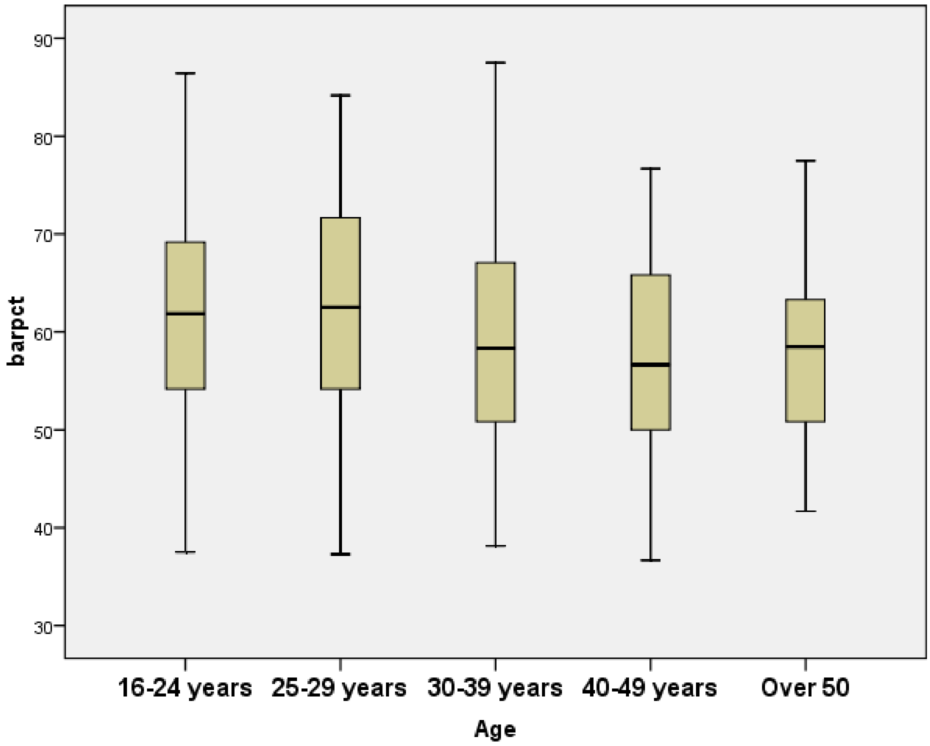


Table 4: Comparison of Test Results by Age

	Mean	Standard Deviation	Sample
Under 25	-0.12	0.84	647
25-29	-.05	0.95	173
30-39	-.21	0.93	107
40-49	-.44	0.76	42
50 and over	-.28	0.74	17
Difference between 20s and 40s	-.39 (raw)	0.43 (standardised)	

Scores by Ethnic Group

Table 5a shows larger differences by ethnic group with three quarters of a standard deviation difference in test scores between White candidates and others. This result is typical of other findings where difference of one standard deviation or more are common. Table 5b shows the results when comparing Degree grades (where a first is awarded 10 points, a 2.1 8 points and a 2.2 6 points). Unfortunately the group sizes are much smaller but there are similar differences with half a standard deviation difference in the black white comparison and 0.88 for the larger Asian-White comparison. These differences are likely to be reflected in success rate on the course. This will be explored in the later phases of the analysis. Differences of this order are likely to lead to differences in success rates for the groups. These will only be justifiable where there is a strong relationship between the test score and performance.

Figure 5 shows that although there are substantial average differences there are low and high scorers in each group.

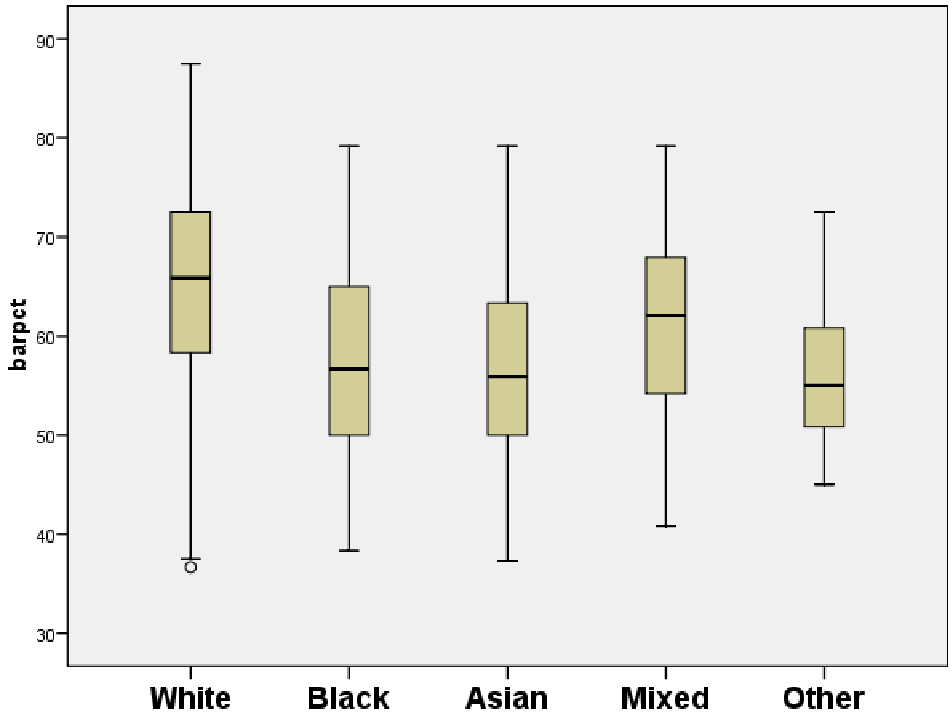
Table 5a: Comparison of Test Results by Ethnic Origin

	Mean	Standard Deviation	Sample
White	0.14	0.85	533
Black	-.49	0.78	82
Asian	-.52	0.72	265
Mixed	-.11	0.88	48
Other	-.51	0.70	35
White-Black Difference	0.63 (raw)	0.75 (standardised)	615
White-Asian Difference	0.66 (raw)	0.82 (standardised)	798

Table 5b: Comparison of Degree Grades by Ethnic Origin

	Mean	Standard Deviation	Sample
White	8.3	1.3	93
Black	7.7	.8	7
Asian	7.2	1.0	22
Mixed	8.5	1.0	4
Other	6.0	.0	3
White-Black Difference	0.6 (raw)	0.47 (standardised)	100
White-Asian Difference	1.1 (raw)	0.88 (standardised)	115

Figure 5. Score by Ethnic Group



Scores by Disability

75 students reported that they had some kind of disability. A disability could impact on test scores in some cases. Table 6 shows the breakdown of scores by disability. These results should be treated with caution as the numbers in each group are quite small and therefore be subject to some variability. Overall those with a disability have a similar level of performance on the test than those without. The largest group are those with dyslexia and their performance is on average a little better than that for the group that did not have any disability. These results suggest there is no prima facia evidence that candidates with disabilities would be disadvantaged by having to take the test although the overall result could mask difficulties for individual students.

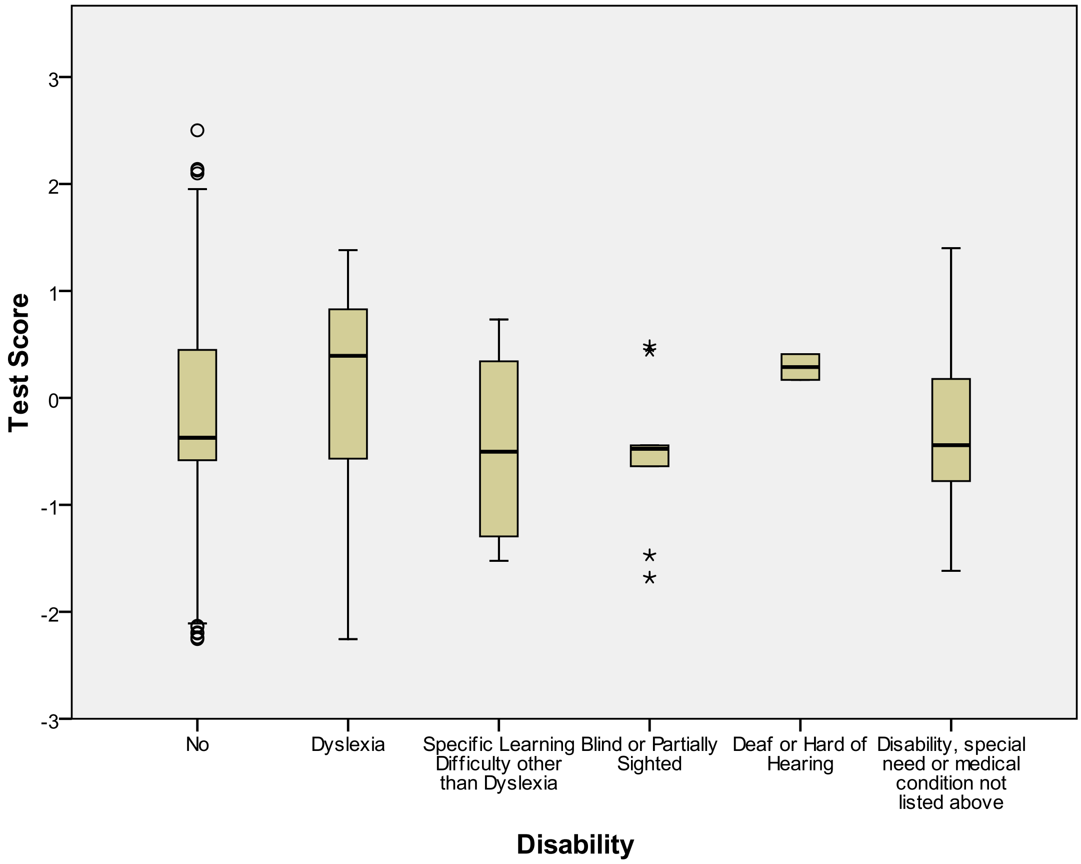
It is recommended that the experiences of these disabled students be followed up to ensure that they received any necessary accommodations to allow them to complete the test effectively.

Table 6: Comparison of Test Results by Ethnic Origin

	Mean	Standard Deviation	Sample
No Disability	-.14	.87	783
Dyslexia	.03	1.02	31
Specific Learning Difficulty other than Dyslexia	-.43	.83	10
Blind or Partially Sighted	-.54	.73	9

Deaf or Hard of Hearing	.29	.17	2
Disability, special need or medical condition not listed above	-.29	.85	23
All with a disability	**-.19**	**0.91**	**75**

Figure 6. Score by Disability



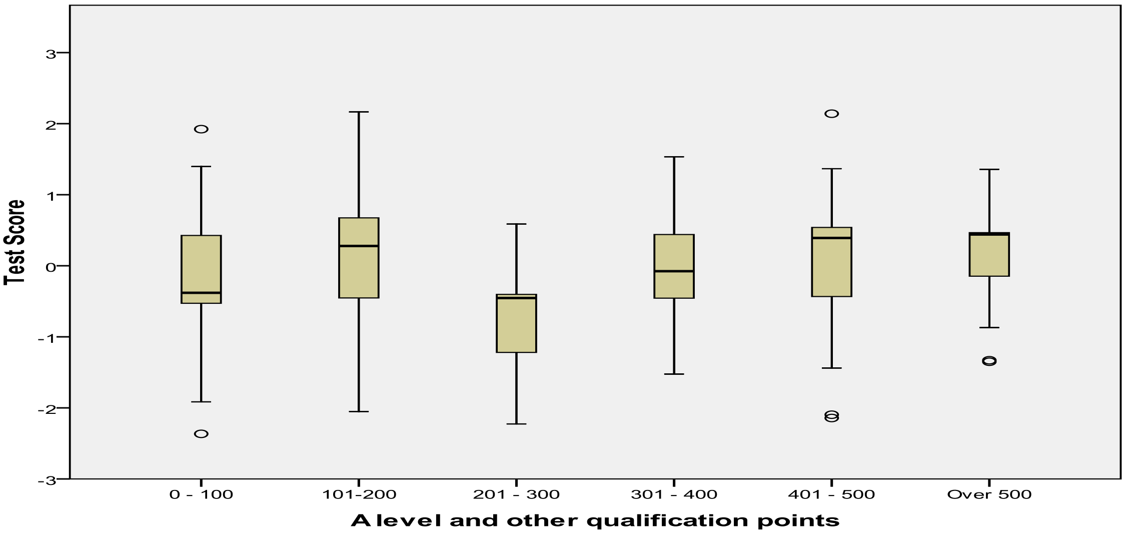
Scores by A Level Results

Secondary external examination results were available for some of the sample and these were translated into UCAS points where possible. UCAS provides guidance on translating results of different examinations into a points scale. Not all students provided sufficient data regarding their school examination results to allow conversion to UCAS points and some may not have provided a complete record and therefore their points allocation may underestimate their performance. The first two groups reported only up to 200 points but this would be insufficient for University entrance without alternative qualifications therefore the results for these two categories are likely to be unreliable. The results below are for the data that was available. There is some inconsistency in the results for the lowest scoring groups but elsewhere there is a strong relationship between the tests scores and A Level results as would be expected.

Table 7: Comparison of Test Results by UCAS A Level Points

	Mean	Standard Deviation	Sample
0 - 100	-0.24	0.78	144
101-200	0.11	0.95	119
201-300	-0.59	0.67	31
301-400	-0.10	0.81	57
401-500	0.16	0.84	85
Over 500	0.24	0.68	31
Difference between Under 300 and over 500	-.83 (raw)	1.23 (standardised)	467

Figure 7: Test Score by A Level Points



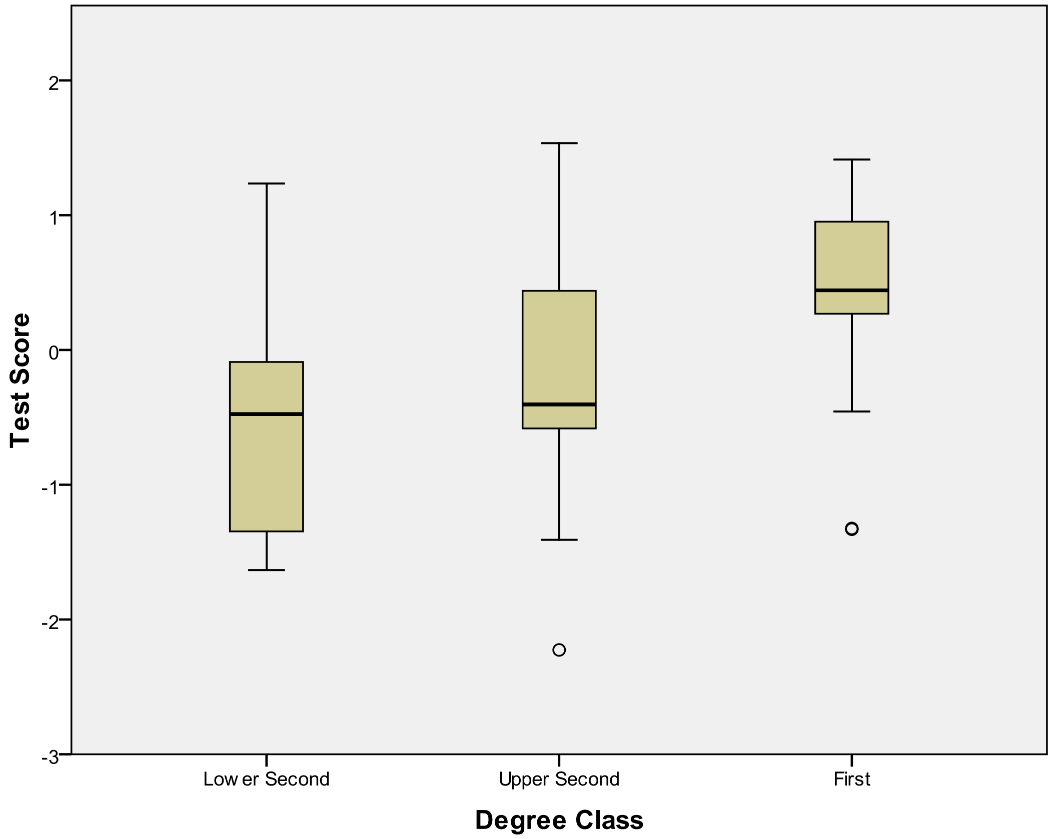
Scores by Degree Class

Only a small proportion of the sample responded to the question regarding their undergraduate Degree class. The following table and figure show the relationship with test score. As with other educational results there is a strong relationship.

Table 8: Comparison of Test Results by Undergraduate Degree Class

	Mean	Standard Deviation	Sample
Lower Second	-.56	.78	36
Upper Second	-.22	.78	111
First	.41	.73	39
Difference between First and Lower Second	-.97 (raw)	1.28(standardised)	186

Figure 8: Test Score by Undergraduate Degree Class



Scores by Place of Undergraduate Degree

Only 370 of the students gave the institution for their first Degree. In addition many institutions were represented by a single student. The following table shows the results broken down by place of first Degree for those institutions with at least 5 respondents. There is a range of results with students from Oxford having the highest overall score and those from newer universities tending to have lower average scores.

Table 9: Comparison of Test Results by place of Undergraduate Degree

University	Mean	Std. Deviation	N
University of Oxford	.77	.55	23
University of Cambridge	.28	.88	13
BPP Law School	.24	.61	6
University of Exeter	.22	.36	5
University of West England	.08	.73	5
Open University	.07	.68	6
City University	-.04	.68	26
University of Manchester	-.12	.73	7
University of Leeds	-.22	1.00	8
College of Law	-.23	1.01	12
De Montfort University	-.45	.02	5

Manchester Metropolitan University	-.45	.71	7
Northumbria University	-.50	.64	14
Nottingham Trent University	-.61	.50	6
University of Liverpool	-.68	.66	5
Total	**-.17**	**.88**	**1501**

Conclusions

The test scores provide a good distribution of scores which will support their use in selection to the BPTC.

There are some differences in scores between groups but these are tending to be at the lower end of the range typical of such comparisons. Further investigation is required in the analysis of the validation data for the test.

The Evaluation of the Aptitude Test for the Bar Professional Training Course (Second Pilot)

Report 2: Initial Validation

September 2011

1. Background

The Bar Standards Board is currently piloting a critical reasoning test for use as part of the recruitment procedure for the Bar Professional Training Course. It is important that before the test is implemented all the appropriate research and checks are carried out to ensure that the test is effective, appropriate, fit for purpose and is fair to all candidates.

A small pilot was undertaken with the 2009/10 student cohort which suggested that the proposed test had the potential to be an effective predictor of course outcomes. Good prediction is necessary for the identification of candidates who are unlikely to have the capacity to complete the course successfully. The pilot also identified some differences in performance between groups.

In order to further evaluate the use of the test, a larger pilot was undertaken with the 2010/11 cohort. The first report in this series reviewed the test score distributions and compared the scores for different groups. This report provides the results relating to the prediction of course outcomes based on the results of first sit examinations. A further report will deal with prediction of course outcomes including Autumn re sit results and use of the test in practice.

2. Data

2.1 Sample

All current students on the BPTC were asked to attend a Pearson Vue testing centre to take the test during November 2010. The students were asked to take the test relatively early in their studies to minimise the impact of the course on their performance.

1501 students attended the centres and completed a version of the test.

The colleges were asked to report the grade and final examination result for students after the first examination sittings in summer 2011. Grade data was supplied for 1370 individuals and a final examination result for 728 people.

Just fewer than 1000 students completed the demographic information requested. Table 2.1 shows the demographic make up of the different parts of the sample from the data available. The samples are very similar but there is a small trend for results to be more likely to be available for younger white students.

Table 2.1: Demographic Details

	All taking test	With grade information	With examination score
% Female	55 %	56%	57%
% White	56 %	55%	61%
% English Primary Language	92 %	92 %	92 %
% Aged under 25	66 %	70 %	70%
% Have a disability	9 %	8 %	8 %
Approximate* Sample with data	994	844	461

* numbers differ slightly for each demographic indicator

Table 2.2 shows the breakdown of the sample by college of study. All institutions were able to provide information on grades for at least some of their students but only 4 provided examination results. However these included the largest course providers.

Table 2.2: College of Study

	Percent All	Percent of students with Grades	Percent of students with Examination scores
BPP	22.5%	20.6%	35.7%
City	19.7%	21.8%	25.7%
College of Law	18.2%	19.8%	34.5%
MMU	7.8%	7.9%	
UNN	7.8%	8.3%	
UWE	6.7%	7.8%	
Nottingham	4.9%	5.7%	4%
Cardiff	4.1%	4.4%	
Kaplan	3.2%	3.8%	
Not identified	5.1%	0%	0%
Total (100%)	**1501**	**1270**	**728**

2.2 Test Score

For the current test students completed one of 5 versions of the test which were each longer than the test for operational use. The test versions contained a mixture of new items and those used in the previous trial. This was to enable the piloting of a large pool of questions which would support multiple test versions in operational use. In order to accurately compare across the different test versions the item pool was calibrated using a 3 parameter logistic item response theory model with fixed guessing. The score has a scale with a mean near zero and a standard deviation of

1. This is an arbitrary scale and in reporting scores this can be transformed into one which will be easier to interpret on an individual score basis.

2.3 Course Results

Two criteria were used to reflect performance on the course. The first is the course grade outcome. Depending on examination results students are graded as Outstanding, Very Competent, Competent or Not competent. The last is a failing grade. In addition, since these are interim results before all re sits are completed some candidates may be ‘referred’ to re sit examinations in the Autumn. A handful of students were deferred for some reason and had no results. There were also some candidates on part time courses who had year one results rather than finals.

The main purpose of the test is to identify in advance applicants likely to fail. The current results have only limited power in identifying potential failure since most of those who have failed one or more elements will be referred to re sit these papers in September and some will go on to gain a passing grade. It is not possible to differentiate those re sitting papers because of poor performance at the first sit from those scheduled to re sit due to personal circumstances.

The second criterion is average examination result. Examination results are expressed as percentages and the average across all papers sat by the student is the criterion used here. Examination results have a strong relationship with grades in that a certain level of examination performance is required for each grade but they provide a more differentiated indicator of performance.

It is not possible to identify those students who took the test in November 2010 but later dropped out of the course before completion. There is no grade for 231 students who took the test originally but it is not possible to currently to identify which are students who have since dropped out of the course and which are ones for whom scores have not been matched due to administrative issues. It is likely that this group contains a proportion of people who struggled with meeting the course requirements who might have been identified by the test. It would be useful to try to identify students who drop out for the final stage of the analysis.

Table 2.3 shows the distribution of test results and examination scores for the sample. The means on the test for those with criterion scores are a little higher (less negative) than for those without. The trend does not reach statistical significance with respect to the availability of grades but those for whom examination scores have been reported have significantly higher test scores than those for whom this information is not available. This suggests a trend for weaker students to be more likely to be missing examination results. This is unfortunate since it is the weaker students who are the main focus of this analysis. It is to be hoped that the phase 3 analysis which includes the re sit data will be based on a more complete sample.

Figure 2.1 shows the distribution of examination results. Most are scoring over 60 with only a small proportion having lower scores. This reflects the fact that many of the lower scorers may have been referred and a final examination grade not provided.

Table 2.3 Distribution of test results and examination scores

	Mean	**Standard Deviation**	**N**
Test Score All	-0.17	0.88	1501
Test Score for those with Grades	-0.15	0.87	1235
Test Score for those with Examination Score	-0.01	0.87	728
Test Score for those with no outcome information	-0.24	0.90	231
Examination Score	73.75	8.12	728

Figure 2.1: Distribution of examination results

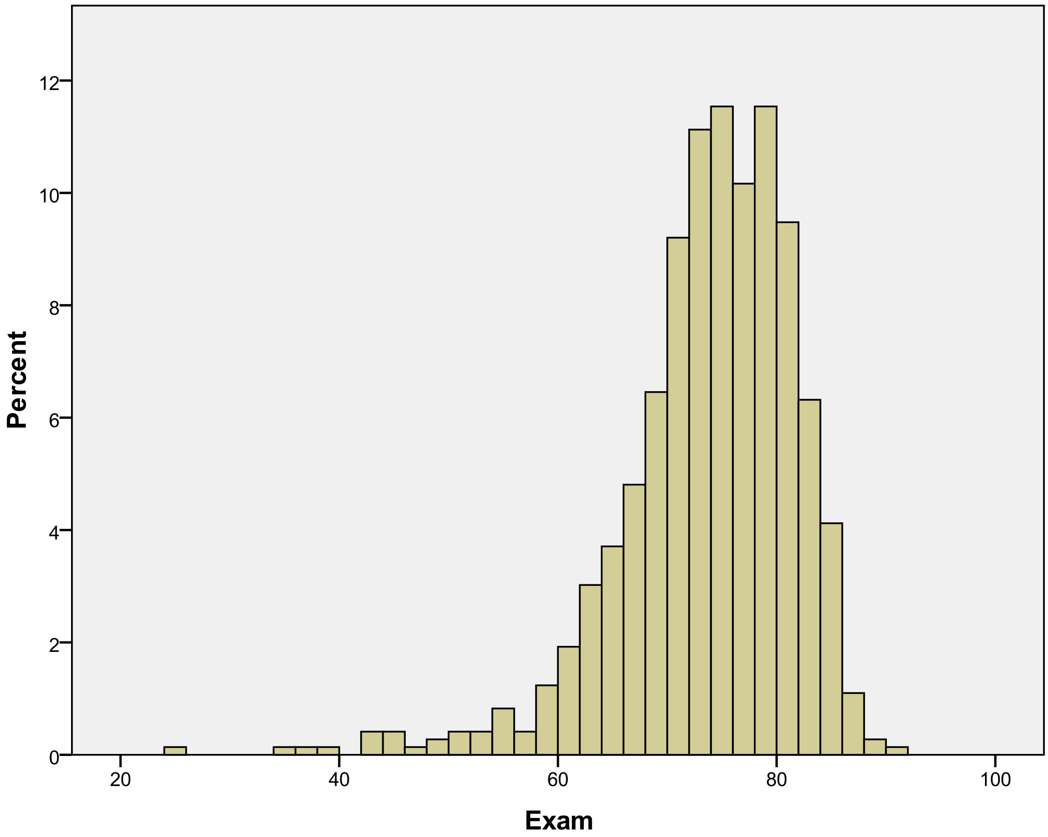


Table 2.4 shows the distribution of grades. Just over forty percent were rated very competent and another forty percent have no final grade pending re sits in

September or for some other reason. Nearly 10% are rated Outstanding. At this stage only 3% have received a failing grade. However it is expected that a proportion of those referred for re sits will also fail. This means that the current sample does not represent failing students well and it is this group that the test is intended to identify. Figure 2.2 shows the distribution in a pie chart.

Table 2.4: Distribution of Grades

	Frequency	Percent
Outstanding	114	9.1%
Very Competent	522	41.5%
Competent	51	4.1%
Pass year one	13	1.0%
Referred	506	40.2%
Referred year one	10	0.8%
Fail	42	3.3%
Total	1258	100.0%
Deferred	12	
Other Missing	231	
Total	**1501**	

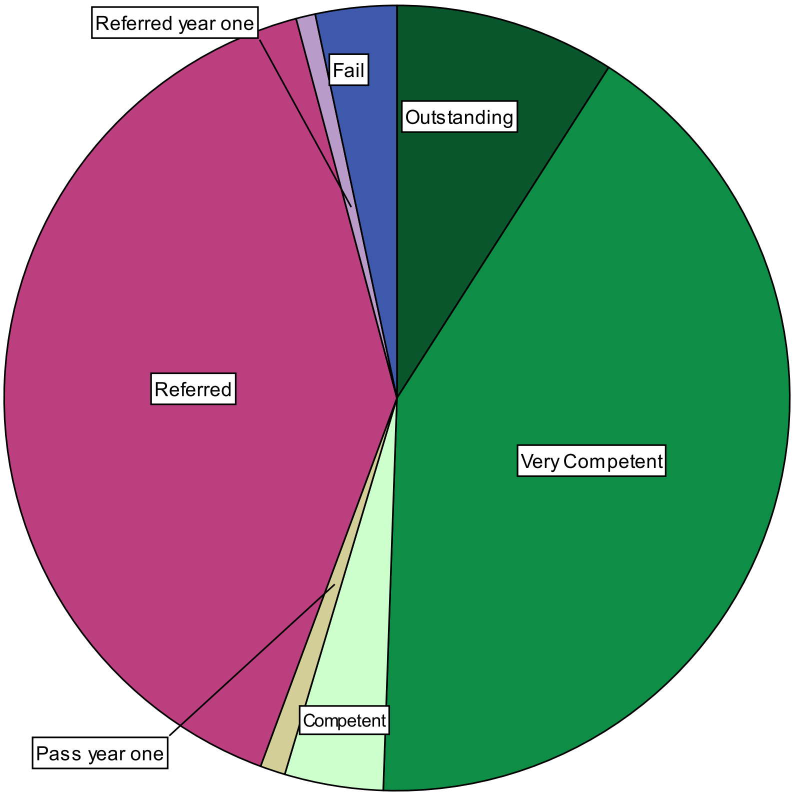


Figure **2.2: Distribution of course grades in sample (Deferred omitted)**

Table 2.5 shows the proportion of students at each grade for whom an examination result was and was not supplied. This shows that examination results were more likely to be provided for those with higher grade outcomes and least likely to be provided for those who failed or were referred. The implication is that analyses with examination scores will under represent poorer performers.

Table 2.5: Grade distribution by examination score inclusion

	Examination Score Provided	**Examination Score not Provided**	**N**
Outstanding	75%	25%	114
Very Competent	72%	28%	522
Competent	65%	35%	51
Pass year one	0	100%	13
Referred	43%	57%	506
Referred year one	0	100%	10
Fail	33%	67%	42
Deferred	0	100%	10
Total	57%	43%	1270

3. Prediction of Course Outcomes

3.1 Predicting Examination Results

A good relationship between test scores and course outcomes is critical to using the test to identify people unlikely to pass the course. A correlation is a statistic which provides an estimate of the size of the relationship between two variables. A correlation ranges from -1 to 1 where 1 is a perfect linear relationship and -1 is a perfect negative relationship (i.e. as one variable rises the other falls). In this case a positive relationship between examination results and test scores is desirable. A correlation of 0.3 or above is desirable in using a test for selection although even values of 0.2 can be useful.

The correlation between examination results and test score for the 728 students for whom data is available was 0.49. This value is highly statistically significant and shows that there is a strong relationship between the test and examination performance. Because it is likely that poorer students are less likely to have had an examination result reported in this sample, the value may underestimate the true relationship.

In order to use the test score to predict outcomes the relationship needs to be expressed as a function. A regression analysis was used to identify the optimal function. This is an interim analysis. Because many of the students with low examination results are not represented in the sample the result will have some unreliability. It is likely to underestimate the strength of the relationship between the test scores and examination results. In the final phase a cross validation approach will be taken. This is not performed at this stage since it would overestimate the robustness of the findings as it cannot take into account the bias in the sample.

A linear regression was performed for the 728 students with both test scores and examination results. Entering just the single variable for test score into the equation results in a multiple R identical to the correlation. The multiple R is an indicator similar to a correlation which can express the relationship between a linear combination of variables and another variable. Where only one variable is used to predict the value of another, the multiple R will be identical to the correlation. The squared multiple R is an estimate of the proportion of variance explained by the predictor variable. In this case the value is 0.24 – 24% of the variation in examination results can be explained by the test. The result is highly statistically significant.

Non-linear regression was explored with quadratic, cubic, logistic and logarithmic equations. However the non-linear solutions were no better than the linear result. The simpler linear model is preferred where alternative models do not improve fit to the data.

Table 3.1 shows the standardised and unstandardised coefficients. The unstandardised coefficients represent the multiplier and constant for a linear equation to predict examination results from test score. The standardised coefficient expresses the multiplier in standard deviation units and provides results that are more comparable across variables when more than one predictor is used. The standardised coefficient for test score is just under a half. This means that the examination score is predicted to rise by about half a standard deviation for every standard deviation rise in the test score.

Table 3.1: Regression of Examination results on Test Score

Predictor	**Unstandardised coefficient**	**Standard Error**	**Standardised Coefficient**	**Statistical significance**
Test Score	4.58	0.30	0.493	p<0.001
Constant	73.81	0.26		p<0.001

The unstandardised coefficients provide a linear equation to convert test scores into predicted examination scores. Table 3.2 shows some example results of applying the prediction equation. The prediction is the best estimate of the future examination score but it will not be exact. There are many factors that affect examination scores apart from ability as measured by the test. The amount of work the student does will have a large impact along with such factors as previous understanding and knowledge, personal situation and interest in the course.

The standard error of estimate (SEE) is the typical difference between the predicted examination score and the actual score achieved. The standard error of estimate in this case is 7.1. This means for any student there is a 67% chance that their actual examination score will lie between one SEE above the predicted score and one SEE below. There is a 96% probability that the actual examination score will be within 2 SEE of the predicted score. Table 3.2 shows the 96% confidence interval for scores.

If the examination pass score is 60 then a test score of -3 or above would predict this result. While the pass score for each module is 60, students with an average examination score of 60 probably fail since they will have some scores below 60 as well as some above. Again the absence of data on the lower performing students makes it difficult to estimate what the lowest desirable predicted average examination score should be.

Table 3.2: Predicted examination scores from test scores

Test Score	**Predicted Examination**	**96% confidence interval**
-3	60.1	46-74
-2	64.5	41-79
-1	69.2	55-83
0	73.8	60-88
1	78.4	64-93
2	83.0	69-97
3	87.5	73-100

3.2 Predicting Grades

In order to explore the relationship between the test score and the examination grades the test score was split into 5 bands each encompassing approximately 20% of the sample. Table 3.3 shows the comparison between these bands and the grade outcomes. Cells containing more than 20% of a grade band are shaded. This shows that 77% of those who were graded outstanding were in the top 2 score bands on the test whereas 71% of those who failed were in the bottom two score bands.

Table 3.3: Relationship between test scores and grades

	Grade							
Test score range	Outstanding	Very Competent	Competent	Pass year one	Referred	Referred year one	Fail	Total
1 Low	7.0%	10.2%	11.8%	15.4%	30.2%	70.0%	35.7%	19.4%
2	6.1%	15.3%	33.3%	30.8%	23.7%	10.0%	35.7%	19.4%
3	9.6%	23.2%	23.5%		21.5%	10.0%	19.0%	20.8%
4	22.8%	26.1%	15.7%	15.4%	15.2%	10.0%	9.5%	20.2%

5 High	54.4%	25.3%	15.7%	38.5%	9.3%			20.2%
Total (100%)	114	522	51	13	506	10	42	1258

Table shows Percentage of each grade scoring within score range

Considering the breakdown in the other direction, of those scoring in the highest band on the test, 24% are graded outstanding compared to only 9% overall, only 19% are referred and none fail. In contrast 66% of those in the lowest band on the test are referred and a further 6% fail. Only 28% pass at first sitting.

The relationship is not perfect. There is a small percentage of those graded Outstanding or Very Competent that are in the lowest 2 bands of the test. Some of these may be students who did not invest any effort in completing the test and therefore have scores which do not reflect their ability. However the prediction from the test will not be perfect and although these results show there will be a substantial improvement in outcomes if the test was used in selection, its use will exclude students who would have done well on the course. Of course this will be true of any selection rule. There will be some people with a third class Degree who might have done well on the course. Like those with low test scores, they will be exceptions to the general rule that those with better qualifications tend to do better on the course. Section 5 provides results regarding the effectiveness of applying a cut based on the test scores.

4. Incremental Validity

The previous section shows that the test score is predictive of examination results. However it may be the case that other information could do this equally as well. In the current data set there is information about A Level (or equivalent) results and first Degree grades that could provide similar information. This information is only available for part of the sample. Using this data a two phase multiple regression analysis was performed to look at the incremental prediction of the test above these factors. A similar approach could be undertaken with IELTS scores but these are not available in the current data set.

Table 4.1 shows the means, standard deviations and samples available for each of the variables for this analysis. As discussed earlier the A Level points data may have low reliability because many students had only a single A Level result recorded which would be insufficient for university entrance without other qualifications.

Table 4.1: Mean and standard deviation for regression variables

	Mean	**SD**	**N**	**Correlation with Examination result**
A Level Points	255	181	238	0.19**

Degree Class	8.0	1.3	186	0.38**
Examination result	74.2	8.1	238	n/a
Test Score	0.04	0.85	238	0.49**

**p<0.01

The most appropriate regression model evaluates the impact of A Level points and Degree class in predicting results and then the incremental value of the test scores above this. However because only a small part of the sample had data for all the relevant variables additional analyses were run considering Degree class and A Level points separately for which larger samples were available. The results are provided in table 4.2

Model 1 is the original model with just the test scores used as a predictor and these results repeat those presented earlier. Model 2 looks at the value of the test after A Level points have been taken into account. Model 3 considers the test value after Degree class is taken into account. The last model looks at the value of the test after both Degree class and A Level points are considered.

The standardised regression coefficients (beta) for each variable are provided as these provide an indication of the relative importance of the different factors. In all the models the test score has the highest beta coefficient which is always statistically significant. A Level points also have significant coefficients. Degree class reaches significance without A Level points but when these are included it does not reach significance although this analysis is performed on a smaller sample. Overall the test score is shown as the strongest single predictor and Degree class as the weakest predictor of examination results.

The Multiple R and Adjusted R-Square increase quite substantially as variables are added to the equation. This conclusion is that including educational information together with the test will provide better prediction than the test alone. As these values increase the Standard Error of Estimate reduces. This shows how the accuracy of the predicted examination result increases with the inclusion of more variables in the model.

The incremental R-squared for the test scores is large and always statistically significant. This shows that the test improves the prediction of examination scores by a substantial amount even when educational qualifications are taken into account.

Table 4.2: Comparison of Models to predict Examination score

Predictors	**Model 1**	**Model 2**	**Model 3**	**Model 4**
	Test Only	**A Level + Test**	**Degree + Test**	**A Level, Degree + Test**
Sample Size	728	238	96	77

A Level Points Beta		0.142*		0.236*
Degree Class Beta			0.186*	0.093
Test Score Beta	0.493**	0.512**	0.473**	0.443**
Multiple R	0.493**	0.543**	0.576**	0.630**
Adjusted R-squared	0.242**	0.289**	0.318**	0.372**
Incremental R-squared for test	0.243**	0.260**	0.186**	0.147**
Standard Error of Estimate	7.07	6.81	5.77	5.60

* p<0.05; **p<0.01

5. Selecting a Cut Score

The results presented in this section, together with those from the previous section can inform cut score decisions. The following table shows the impact of a range of cut scores on the proportion of students passing, failing or being referred after the first sit examinations. In the current data set only 42 students are identified as failing. This is too small a number to accurately identify a cut score. In the next phase it will be possible to revisit this analysis once the re sit results are known which will identify a larger group who fail the course.

Table 5.1 shows the impact of applying different cut scores to the current sample. Part time students have been excluded from this analysis leaving a total of 1235 students with identified grades. The bottom row shows that by selecting the best 72% of the sample based on their test scores (Cut score -0.5) the failure rate can be reduced by 30% (from 3.4% down to 2.3%), the referral rate by 20% (from 41% down to 33%) and the pass rate increases by 17% from 56% to 65%). Alternative results for lower (less selective) cut scores are also shown.

Table 5.1: Impact of Cut Score on pass rates

Cut Score	% Students Selected	% Passing	% Referred	% Failing
No Cut Score	100%	56%	41%	3.4%
Cut score -1.5	97%	57%	40%	3.3%
Cut score -1.25	83%	61%	36%	2.8%
Cut score -1.0	79%	62%	35%	2.6%
Cut score -0.75	77%	63%	34%	2.5%
Cut score -0.5	72%	65%	33%	2.3%

Table 5.2 provides a more detailed breakdown of the impact of a cut score of -0.5. Of those who fail the examination on 50% would have passed the test with this cut

score. For those referred at this stage, just under half (42%) would have failed the test whereas only 10% of those who were graded outstanding would not have passed the test. It should be remembered that in this sample some people may have significantly underperformed on the test due to lack of motivation.

Table 5.2: Predicted test success rates at each grade with a cut score of -0.5

Grade	**Fail Test**	**Pass Test**	**Total (100%)**
Outstanding	10%	90%	114
Very Competent	16%	84%	522
Competent	24%	76%	51
Referred	42%	58%	506
Fail	50%	50%	42

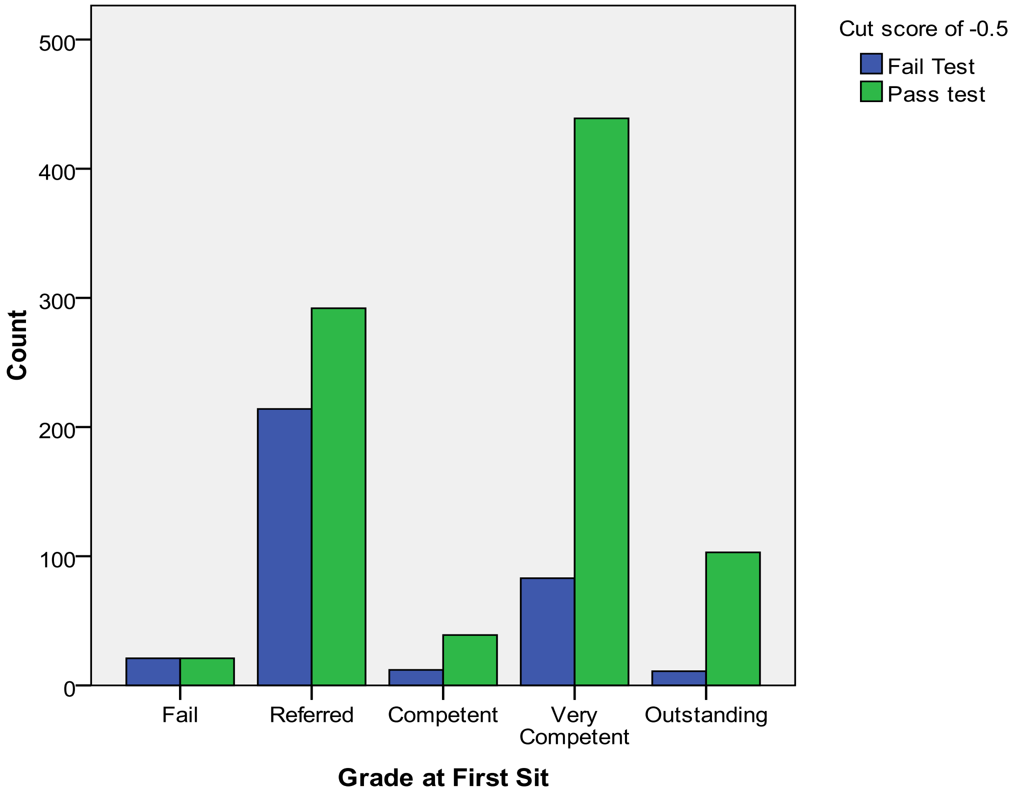
Table 5.3 shows the proportion of students achieving each outcome grade for the full sample and for those that would have been selected applying a cut score of -0.5. The table shows that with the cut score applied there is an increase in those attaining higher grade outcomes and a reduction in the percentage failing or being referred.

Table 5.3 Percentage at each grade with and without cut score of -0.5

Grade	No cut score	With cut score
Outstanding	9%	11%
Very Competent	41%	48%
Competent	4%	4%
Referred	40%	32%
Fail	3%	2%

Figure 5.1 compares the distribution of grades for the unselected and selected group with a cut score of 0.5. It shows that the majority of those who would have failed the cut score are referred at this stage.

Figure 5.1: Distribution of grades with a cut score of -0.5



It is clear from these results that success rates and proportions achieving higher grade passes can be increased by the use of the test. It should be remembered that this sample is pre-selected using existing selection processes so the impact is incremental over current methods of selecting candidates. Even a low cut score on the test results in some improvement. However the impact of the test is probabilistic and there will always be some incorrect decisions.

Because of the large proportion of the sample referred for re sits it is inappropriate to determine a final cut score on the basis of these results. However the tables presented here show how a cut score might work and raises the policy issues that need to be decided such as the relative value of false positive (selecting a candidate who later goes on to fail) and false negative (rejecting a candidate who would have completed the course successfully) decisions, and the overall proportion to be rejected by the test.

6. Group Comparisons

The monitoring report provided a breakdown of test scores by some of the demographic variables before the examination results were available. The analysis showed small differences for most comparisons but larger differences with respect to Ethnic Group. Differences in test scores will lead to adverse impact when the test is used in selection. Such impact can only be justified if it can be shown that the differences are related to course outcomes for that group. The statistical approach to looking at this is a hierarchical regression somewhat similar to the incremental approach used in section 4. If including the group membership variable in the regression equation significantly improves prediction it suggests some bias in the results. Both the group membership variable and its interaction with test score are tested. If there is no improvement in prediction then the score differences on the test are also reflected in the outcome. Below hierarchical analyses are presented for those variables for which the current sample is large enough.

6.1 Regression Results -Sex

Table 6.1 shows the results for sex. The previous report found there is a small, statistically significant difference in scores with men scoring marginally higher than women. The results below show that there is no incremental validity for sex. The adjusted R square is actually smaller than the value for test score alone and the Beta coefficients for the Sex variables are small and are not statistically significant. This analysis is performed on a large enough sample to conclude that any differences in scores between men and women are likely to be reflected in course outcomes.

Table 6.1: Hierarchical Regression - Sex

Predictors	**Model 1**	**Model 2**
Sample Size	461	461
Test Score **Beta**	0.487**	0.398**
Sex **Beta**		-.039
Sex*Test **Beta**		0.087
Multiple R	0.487**	0.489**

Adjusted R-squared	0.236**	0.235**
Incremental R-squared for Sex and Sex*Test	n/a	0.002

* p<0.05; **p<0.01

6.2 Primary Language

The previous report showed a small but statistically significant difference in test score between those who have English as a primary language and those who do not. However examination results were only available for 35 people who reported English as a second language and this is insufficient to perform the analysis. The analysis may be possible with the extended data set following re sits. Table 6.2 shows that there is a larger difference in examination results to that found with the test.

The correlation between the test score and examination was 0.47 (p<0.01) for the group with English as a primary language and 0.42 (p<0.05) for the group with non-primary English. Statistically significant results for both groups suggest that the test is predictive for both groups although results with a larger sample are desirable.

Table 6.2: Comparison of Test Results by primary language

	English Primary Language	**Other Primary Language**
Sample Test	867	80
Test Mean	-0.09	-0.33
Test Standard Deviation	0.86	0.79
Standardised Difference	0.28*	
Sample Examination	405	35
Examination Mean	74.9	70.1
Examination Standard Deviation	7.0	6.4
Standardised Difference	0.69**	

* p<0.05; **p<0.01

6.3 Regression Results - Age

The previous report found a moderate difference between some age groups with the older students performing less well on the test. In the regression analysis the model including the age variable did provide a small but significant improvement in prediction. The third model was tested which did not include the interaction variable. This was just as predictive as the previous model so it can be concluded that only the direct impact of age is statistically significant. Although the effect does reach statistical significance it is actually very small and will have little or no meaningful impact on decision making. However this analysis should be repeated in the post re sit sample to gauge the likely impact on selection decisions.

Table 6.3 Hierarchical Regression - Age

Predictors	**Model 1**	**Model 2**	**Model 3**
Sample Size	456	456	456
Test Score Beta	0.485**	0.404**	0.476**
Age Beta		-0.133**	-0.136**
Age*Test Beta		0.079	
Multiple R	0.485**	0.505**	0.504**
Adjusted R-squared	0.234**	0.250**	0.251**
Incremental R-squared for Age and Age*Test	n/a	0.019**	0.018**

* p<0.05; **p<0.01

6.4 Regression Results – Ethnic Group

Examination results were available for 271 students who described themselves as white and 170 from other ethnic groups including 102 from various Asian groups. The analysis in the previous report showed that there were substantial differences in test scores between the white group and particularly the Asian students.

Table 6.4 shows the results for the White Asian Comparison. There is a significant improvement in prediction by including the main effect of the ethnicity variable and a further increase for the interaction term. When both are in the model the main effect for test score no longer reaches statistical significance. This means that separate regression equations for White and Asian students provides better overall prediction than the combined regression line.

Table 6.4 Hierarchical Regression – White versus Asian

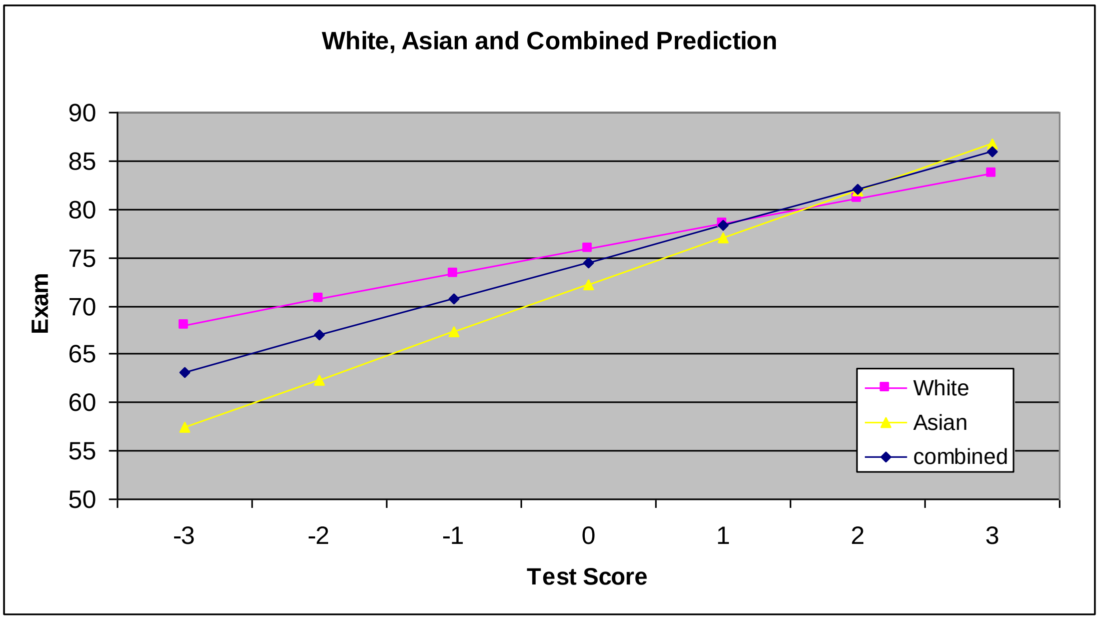
Predictors	**Model 1**	**Model 2**	**Model 3**	**Model 2**
Sample Size	373	373	373	373
Test Score Beta	0.461**	0.042	0.476**	0.042
White/Asian Beta		-0.239**	-0.136**	-0.239**
White/Asian *Test Beta		0.362**		0.362**
Multiple R	0.461**	0.537**	0.526**	0.537**
Adjusted R-squared	0.210**	0.283**	0.273**	0.283**
Incremental R-squared for White/Asian and White/Asian *Test	n/a	0.076**	0.065**	0.012* Increment of Model 2 over Model 3

* p<0.05; **p<0.01

Figure 6.1 shows plots of the different prediction lines. Although the Asian group have a lower average score than the white group, at the same test score, particularly in the lower score ranges, an Asian student would be predicted a lower examination score than a White student with the same test score. This means that using the combined prediction line, shown in blue in the figure, would tend to over-predict the performance of Asian students and under-predict that of white students. This suggests that if there is bias in the test it is against the higher scoring White group rather than the lower scoring Asian group.

Figure 6.1 also shows that the yellow regression line for the Asian group is steeper than the pink line for the White group. The steepness of the slope of the line is an indicator of the predictive power of the test in the sample. The steeper Asian slope means that the test is a better predictor of performance for this group than for the white group. The significant increment in R square of Model 2 with the interaction effect over model 3 without shows that this difference is statistically significant. The test is a better predictor for the Asian group than for the white group although it is a significant predictor for both groups.

Figure 6.1: Comparison of regression equations for separate and combined models



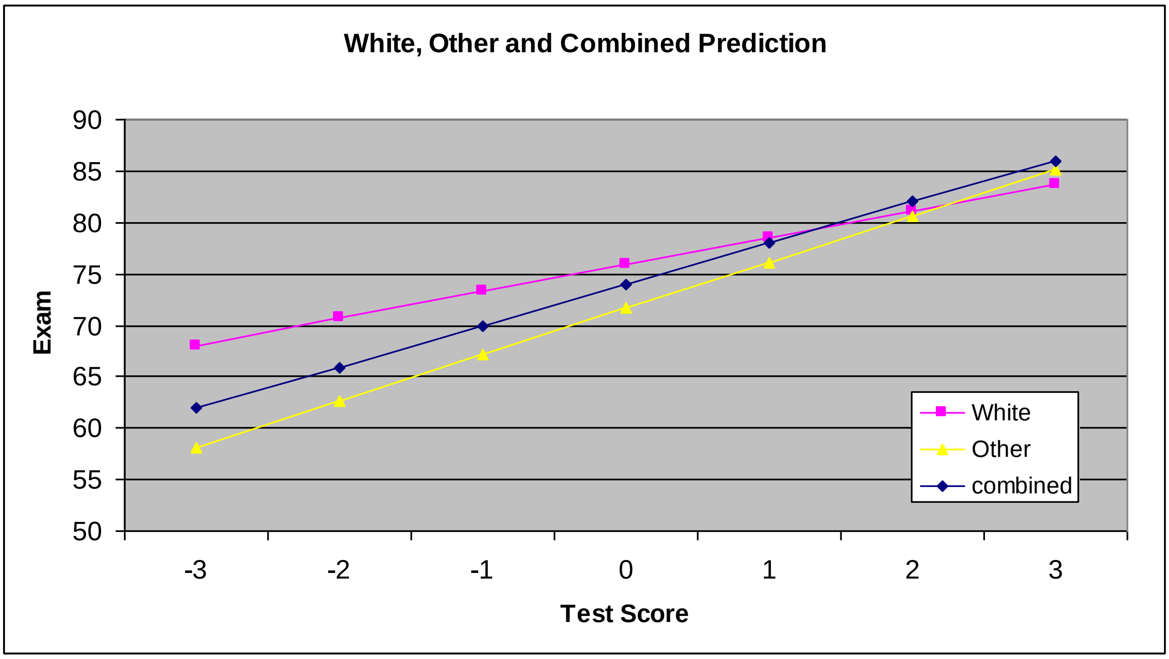
A similar analysis was performed for a comparison between the White group and all other ethnic groups. Table 6.5 shows the regression results. The pattern is very similar to that for the White Asian Comparison. This is not surprising since the majority of the 'Other' group are Asian. Figure 6.2 compares the regression equations and shows that the test is more predictive for the 'other' group and the difference in predictions is most marked at lower score levels where a cut score would be positioned.

Table 6.5 Hierarchical Regression – White versus Other

Predictors	**Model 1**	**Model 2**	**Model 3**	**Model 2**
Sample Size	441	441	441	441
Test Score Beta	0.471**	0.086	0.378**	0.086
White/Other Beta		-0.287**	-0.300**	-0.287**
White/Other *Test Beta		0.311*		0.311*
Multiple R	0.471**	0.559**	0.550**	0.559**
Adjusted R-squared	0.220**	0.308**	0.300**	0.308**
Incremental R-squared for White/Other and White/Other *Test	n/a	0.091**	0.081**	0.010* Increment of Model 2 over Model 3

* p<0.05; **p<0.01

Figure 6.2: Comparison of regression equations for separate and combined models



6.6 Disability

The previous report showed no statistically significant difference in test score between those who have a disability and those who do not. However examination results were only available for 33 people with a variety of different disabilities and this is insufficient to perform the analysis. The analysis may be possible with the extended data set following re sits. The table below shows that there is a larger difference in examination results to that found with the test.

The correlation between the test score and examination was 0.49 for the non disabled group and 0.53 for the disabled group. Both these results reach statistical significance at the 0.01 level. This suggests that the test is predictive for both groups although results with a larger sample are desirable.

Table 6.6: Comparison of Test Results by disability

	No Disability	Disability
Sample Test	783	75
Test Mean	-0.14	-0.19
Test Standard Deviation	0.87	0.91
Standardised Difference	0.06	
Sample Examination	366	35
Examination Mean	74.4	71.7
Examination Standard Deviation	7.0	9.1
Standardised Difference	0.37*	

* p<0.05; **p<0.01

6.6 Regression Results – University type

For this analysis, where it was reported, the university of study for the first or previous Degree was categorised according to whether it belonged to the Russell Group of universities. This was not reported in the monitoring report so the test score comparisons are included here before the regression results are presented.

Table 6.7: Comparison of Test Scores by type of university

	Mean	Standard Deviation	Sample
Russell Group University	.17	.83	92
Other University	-.38	.77	106
Difference	.56 (raw)	.70 (standardised)	198

There is a large difference between the test scores for Russell Group graduates over other university graduates. This is likely to be related to the greater selectivity of the Russell group universities in selecting students. In this data set those who had attended Russell Group institutions had an average of 90 more A Level points than those who had attended other universities.

In the regression analysis the model including the university type variables did not result in a significant improvement in prediction. There is a small improvement but it does not reach statistical significance within the small sample for which this analysis is possible. The implication is that the higher average test scores of Russell Group applicants are on the whole reflected in their performance on the course.

Table 6.8: Hierarchical Regression – Type of University

Predictors	**Model 1**	**Model 2**
Sample Size	101	101
Test Score Beta	0.543**	0.556**
University Type Beta		0.171
University Type *Test Beta		-0.078
Multiple R	0.543**	0.571**
Adjusted R-squared	0.288**	0.305**
Incremental R-squared for University Type and University Type*Test	n/a	0.031

* p<0.05; **p<0.01

7. Example Impact of Cut Score

Where there are score differences between groups this will be reflected in the results of applying a cut score. A smaller proportion of the lower performing group will pass any particular cut. The following tables show the impact of using a cut score of -0.5 on the test on different groups. This is a moderately high cut score with 72% of the sample passing the cut score and 28% failing. The pass rate is consistently higher for those who go on to pass the examination than for those who are referred for re sits or who fail. This means that using the test will result in an increased pass rate as discussed in a previous section.

The tables also show the relative selection ratio for the groups compared. This is the ratio of the pass rate for the group with the lower success rate to that with the higher pass rate. Where this value is below 0.8 the selection fails the ‘four fifths rule’ and is considered to have significant adverse impact. The implication if this is the case in a real selection process is that people from the lower scoring group have less than 80% the chance of people from the higher scoring group of being offered a place to study.

The ratio is above 0.9 for nearly all the comparisons. For university type it is exactly on 80%. For the ethnic group comparison it is just above 0.7. This means that, if the sample here is similar to applicants, people from minority ethnic groups would have just over 70% of the chance that those from the white group would have of passing the test. While this means that if the test is used a smaller proportion of ethnic minority applicants will be offered course places, it is also clear from tables 7.4 and 7.5, even those students who would pass the cut score from the minority ethnic groups have a lower success rate on the course with 80% rather than 90% for the white group passing at the first sitting. This is in line with the regression findings of over-prediction of performance for the lower scoring groups when a common score is used.

Using a lower cut score that more people would pass overall would improve the relative selection ratio for all groups. To illustrate this table 7.6 shows the pass rates for the comparison of White and Asian students with a cut score of -1. The overall pass rate increases from 85% to 89% for the White group but from 61% to 70% for the Asian group. This gives a relative selection ratio of 0.79 which is only just under the 80% level required by the four-fifths rule. A variety of cut scores will be explored when the final outcome data is available. Minimising the Degree of adverse impact should be one of the criteria in choosing a cut score.

Table 7.1: Pass Rate by Gender

Cut = -.5	**Test Pass Rate**	
	Male	**Female**
Sample	365	461
Pass Examination	87%	84%
Referred	65%	58%
Fail Examination	43%	52%
All	77%	73%
Relative Selection ratio	.95	

Table 7.2: Pass Rate by Primary Language

Cut = -.5	**Test Pass Rate**	
	English Primary	**Other Primary**
Sample	724	62
Pass Examination	86%	82%
Referred	63%	63%
Fail Examination	50%	60%
All	76%	73%
Relative Selection ratio	.96	

Table 7.3: Pass Rate by Age

Cut = -.5	**Test Pass Rate**	
	Under 30 years	**30 and over**
Sample	718	101
Pass Examination	85%	84%
Referred	62%	62%
Fail Examination	54%	25%
All	75%	71%
Relative Selection ratio	.95	

Table 7.4: Pass Rate by Ethnic Group

Cut = -.5	**Test Pass Rate**	
	White	**All other**
Sample	435	357
Pass Examination	89%	79%
Referred	77%	51%
Fail Examination	67%	46%
All	85%	62%
Relative Selection ratio	**.73**	

Table 7.5: Pass Rate by Ethnic Group (White Asian)

Cut = -.5	**Test Pass Rate**	
	White	**Asian**
Sample	435	234
Pass Examination	89%	80%
Referred	77%	47%
Fail Examination	67%	38%
All	85%	61%
Relative Selection ratio	**.72**	

Table 7.6: Pass Rate by Ethnic Group (White Asian), Lower Cut

Cut = -1	**Test Pass Rate**	
	White	**Asian**
Sample	435	234
Pass Examination	92%	85%
Referred	80%	61%
Fail Examination	83%	44%
All	89%	70%
Relative Selection ratio	**.79**	

Table 7.7: Pass rate by Disability

Cut = -.5	**Test Pass Rate**	
	Not disabled	**Disabled**
Sample	651	54
Pass Examination	86%	84%
Referred	58%	68%
Fail Examination	48%	0
All	74%	74%
Relative Selection ratio	1	

Table 7.8: Pass Rate by University Type

Cut = -.5	**Test Pass Rate**	
	Russell Group	**Other University**
Sample	81	69
Pass Examination	94%	79%
Referred	65%	60%
Fail Examination	100%	0%
All	85%	68%
Relative Selection ratio	.80	

8. Summary and Conclusions

Course outcome results including examination results and final grades from first sit examinations were collated with test scores and background data.

A strong correlation of 0.49 was found between the examination results and the test scores which shows the test would be a useful selection tool to improve the success rate of selected applicants. A similar strong relationship was found between test score ranges and grade. 54% of students graded outstanding scored in the highest test band with only 7% in the lowest band.

When educational qualifications were also considered it was found that prediction of course outcomes could be improved. However the test score was the best single predictor of examination results and added value when combined with both A Level points and Degree Class or the two combined.

Modelling potential cut scores showed that using the test score to select candidates could improve the proportion of students achieving the highest grade and passing the course at the first sitting. The results of the re sits are required before the potential impact of using the test on the failure rate can be fully gauged.

The predictive power of the test for various subgroups was examined and it was found that, where the sample was sufficient for the analysis, there was good relationship between the test and examination results for all groups. There was a substantial significant finding for the ethnic group comparisons showing that the test was a better predictor for students from minority group and decisions made using the test may tend to favour those from minority ethnic groups even though they perform less well on the test as a whole. A significant but small difference was found in the age analysis.

No substantial adverse impact was found using a moderate cut score for gender, primary language, age or disability. There was a marginal impact for type of university with those attending Russell Group institutions having a higher pass rate. There was a more substantial impact for ethnic group with students from minority groups being less successful than the white group overall.

Because many students of marginal ability were referred for re sits this sample and analysis cannot give a full picture of the ability of the test to identify students likely to fail the course. Further analysis when the re sit results are available and students have received their final grades will be carried out as soon as these results are available.

Recommendations for phase 3 analysis

The phase 3 analysis will be more powerful if further data is available in addition to the post re sit results for all students.

1. Identify, where possible, students who dropped out during the course so that the potential of the test to predict these outcomes can be examined.
2. Provide examination results for all students to avoid the bias in this criterion against lower performers.
3. Provide exact A Level (or equivalent) points data for all students
4. Provide first Degree grade and university of study for all students
5. Try to obtain IELTS scores for those students for whom English is not the primary language

Consider the policy issues surrounding cut scores – such as by how much it is desirable to reduce the failure rate relative to the potential for excluding candidates who might have completed the course successfully.

The Evaluation of the Aptitude Test for the Bar Professional Training Course (Second pilot)

Report 3: Final Validation

November 2011

1. Background

The Bar Standards Board is currently piloting a critical reasoning test for use as part of the recruitment procedure for the Bar Professional Training Course. It is important that before the test is implemented all the appropriate research and checks are carried out to ensure that the test is effective, appropriate, fit for purpose and is fair to all candidates.

A small pilot was undertaken with the 2009/10 student cohort which suggested that the proposed test had the potential to be an effective predictor of course outcomes. Good prediction is necessary for the identification of candidates who are unlikely to have the capacity to complete the course successfully. The pilot also identified some differences in performance between groups.

In order to further evaluate the use of the test, a larger pilot was undertaken with the 2010/11 cohort. The first report in this series reviewed the test score distributions and compared the scores for different groups. The second report provides the results relating to the prediction of course outcomes based on the results of first sit examinations. This report summarises the results from the first two reports and extends the prediction of course outcomes to include the Autumn resit results and use of the test in practice.

2. Data

2.1 Sample

All current students on the BPTC were asked to attend a Pearson Vue testing centre to take the test during November 2010. The students were asked to take the test relatively early in their studies to minimise the impact of the course on their performance. 1501 students attended the centres and completed a version of the test.

The colleges were asked to report the grade and final exam result for students after the first exam sittings in summer 2011 and the second sittings in late summer, early autumn. From the first sittings. grade data was supplied for 1370 individuals and a final exam result for 728 people. Following the resits, grade data was provided for a further 298 students, and an additional 396 received a revised grade. Exam results were provided for 570 students following resits including 256 who had exam results assigned at first sit. 88 existing exam results were revised following resits by 1 point or more.

Just fewer than 1000 students completed the demographic information requested as part of the testing procedure. Table 2.1 shows the demographic make up of the different parts of the sample from the data available. The samples are very similar but there is a small trend for results to be more likely to be available for younger white students.

Table 2.1: Demographic Details

	All taking test	With grade information	With exam score
% Female	55 %	56%	57%
% White	54 %	53%	54%
% English Primary Language	89 %	89 %	89 %
% Aged under 25	66 %	70 %	70%
% Have a disability	9 %	8 %	9 %
Approximate* Sample with data	994	845	640

* numbers differ slightly for each demographic indicator

Table 2.2 shows the breakdown of the sample by college of study. All institutions were able to provide information on grades for at least some of their students but not all provided exam results.

Table 2.2: College of Study

	Percent All	Percent of students with Grades	Percent of students with Exam scores
BPP	22.1%	20.2%	25.9%
City	20.2%	22.3%	27.7%
College of Law	17.9%	19.4%	24.9%
UNN	8.0%	8.5%	0%
MMU	7.7%	7.6%	5.3%
UWE	6.8%	7.8%	3.6%
Nottingham	5.2%	6.0%	7.6%
Cardiff	4.1%	4.5%	2.4%
Kaplan	3.1%	3.7%	2.7%
Not identified	4.9%	0%	0%
Total (100%)	**1568***	**1338**	**1042**

67 of these did not have an identifiable test score

2.2 Test Score

For the current test students completed one of 5 versions of the test which were each longer than the test for operational use. The test versions contained a mixture of new items and those used in the previous trial. This was to enable the piloting of a large pool of questions which would support multiple test versions in operational use. In order to accurately compare across the different test versions the item pool was calibrated by Pearson using a 3 parameter logistic item response theory model with fixed guessing. The scores supplied are on a scale with a mean near zero and a standard deviation of 1. This is an arbitrary scale and in reporting scores this can be transformed into one which will be easier to interpret on an individual score basis. However the reports provided here are based on the raw scale points to ensure that operational recommendations from this report can be implemented accurately on Pearson systems.

Table 2.3 shows the average scores for different parts of the sample.

Table 2.3: Test and Exam Scores

	Mean	**Standard Deviation**	**N**
	Test Scores		
All taking test	-0.17	0.88	1501
First sit grade known	-0.15	0.87	1235
First sit exam score known	-0.01	0.87	728
Grade after resits known	-0.16	0.88	1271
Exam score after resits known	-0.13	0.88	988
No course performance information known	-.24	0.90	230
	Exam Scores		
First Sit Exam Scores	73.8	8.1	728
Exam score after resits	72.0	8.2	1042

2.3 Course Results

Two criteria were used to reflect performance on the course. The first is the course grade outcome. Depending on exam results students are graded as Outstanding, Very Competent, Competent or Not competent. The last is a failing grade. In addition, even after the first resits, some candidates may be ‘referred’ to retake exams. A handful of students were deferred for some reason and had no results. There were also some candidates on part time courses who had year one results rather than finals.

The main purpose of the test is to identify in advance applicants likely to fail. The summer results had only limited power in identifying potential failure since at that stage only 42 students were identified as failing. There were 109 failing students following the resit results.

After the first sit examinations nearly 506 students were referred for resits. Following the autumn resits most of these students received a final grade with only 132 referred for further examinations.

The second criterion is average examination result. Examination results are expressed as percentages and the average across all papers sat by the student is the criterion used here. Exam results have a strong relationship with grades in that a certain level of exam performance is required for each grade but they provide a more differentiated indicator of performance.

It is not possible to identify those students who took the test in November 2010 but later dropped out of the course before completion. There is no grade for 230 students who took the test originally but it is not possible currently to identify which are students who have since dropped out of the course and which are ones for whom scores have not been matched due to administrative issues. It is likely that this group contains a proportion of people who struggled with meeting the course requirements who might have been identified by the test.

Figures 2.1 and 2.2 shows the distribution of exam results. Most students are scoring over 60 with only a small proportion having lower scores. There are more lower scores in figure 2.2 after the autumn resits.

Table 2.4 shows the distribution of grades. Following resits over 50% of the sample attain a grade of Very Competent. Over half the remainder are rated Competent. 8% fail the course and a similar proportion is graded outstanding. Many of the 40% who were referred at first sit have now been given a final grade.

Figure 2.1: Distribution of exam results, first sit

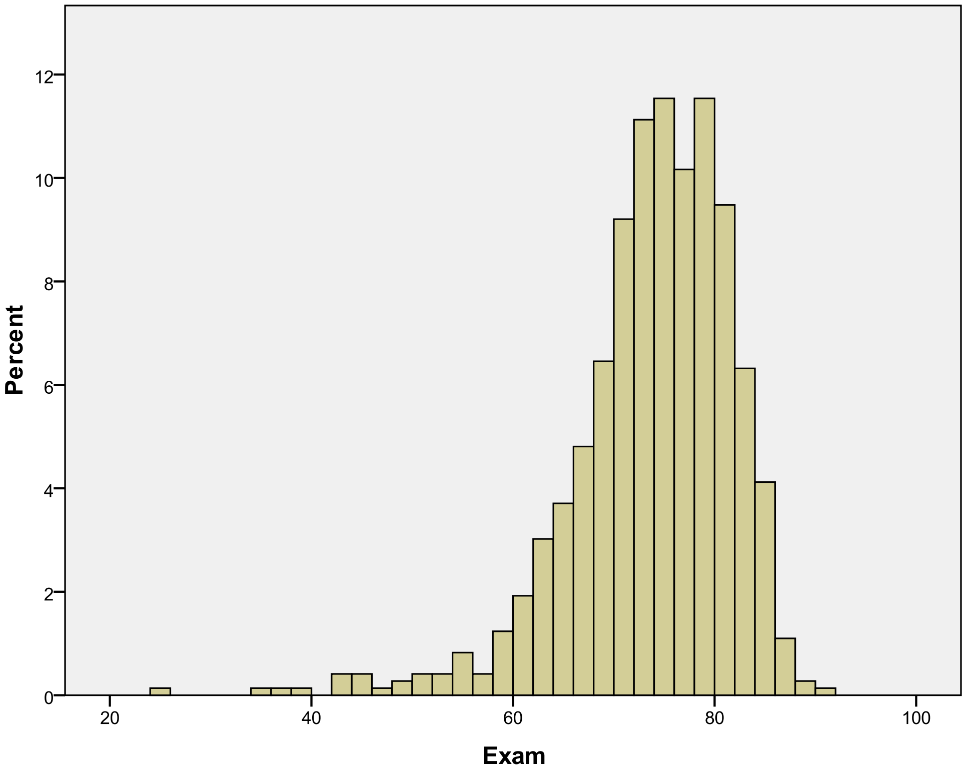


Figure 2.2: Distribution of exam results, after resits

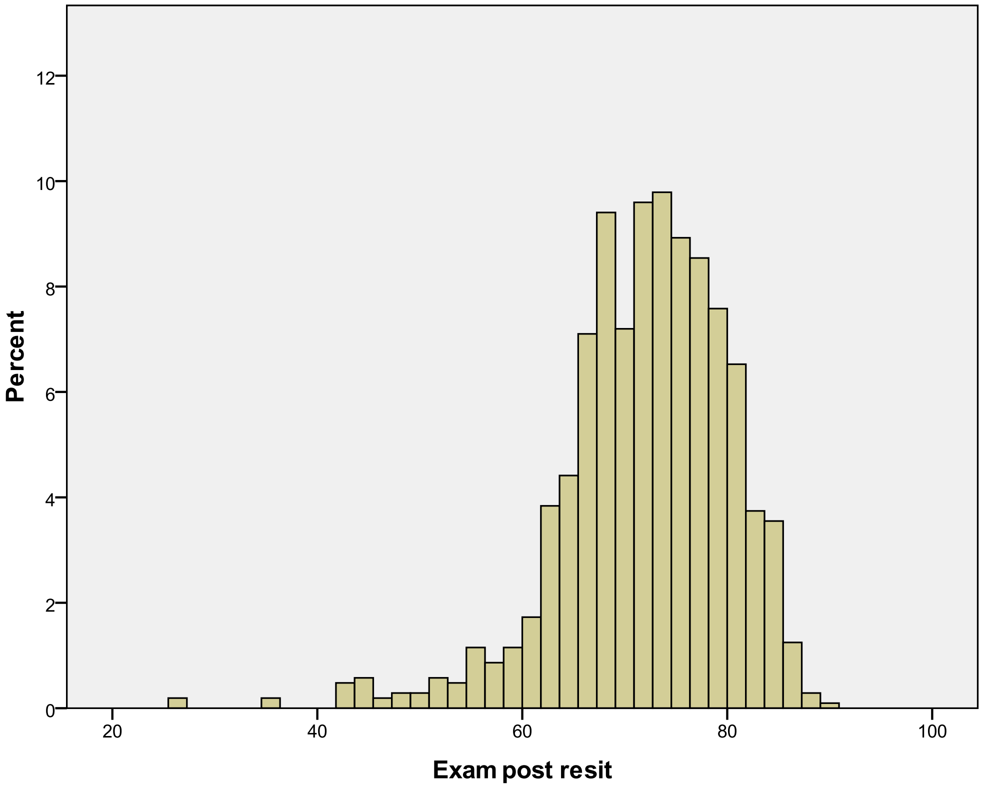


Table 2.4: Distribution of Grades

	First Sit		**After Resit**	
	Frequency	Percent	Frequency	Percent
Outstanding	114	9.1%	117	9.0%
Very Competent	522	41.5%	678	52.2%
Competent	51	4.1%	236	28.3%
Pass year one	13	1.0%	13	1.0%
Referred	506	40.2%	158	11.8%
Referred year one	10	0.8%	12	0.9%
Fail	42	3.3%	110	8.2%
Total	1258	100.0%	1324	100.0%
Deferred	12		14	
Other Missing	231		163	
Total	**1501**		**1501**	

Table 2.5 shows the proportion of students at each grade for whom an exam result was supplied before and after resits. This shows that following resits, exam results were more likely to be provided for all groups. The bias towards providing exam results for those with higher grade outcomes evident with the first sit results has disappeared. Only for those referred and deferred is the percentage of results available below 75% and this may reflect a genuine absence of exam results for some students in these groups.

Table 2.5: Grade distribution by exam score inclusion

	Exam Score Provided after first sit	**Exam Score Provided after resits**
Outstanding	75%	75%
Very Competent	72%	79%
Competent	65%	88%
Pass year one	0%	0%
Referred	43%	66%
Referred year one	0%	0%
Fail	33%	92%
Deferred	0	50%
Total	57%	78%

3. Prediction of Course Outcomes

3.1 Predicting Exam Results

A good relationship between test scores and course outcomes is critical to using the test to identify people unlikely to pass the course. A correlation is a statistic which provides an estimate of the size of the relationship between two variables. A correlation ranges from -1 to 1 where 1 is a perfect positive linear relationship and -1 is a perfect negative relationship (i.e. as one variable rises the other falls). In this case a positive relationship between exam results and test scores is desirable. A correlation of 0.3 or above is desirable in using a test for selection although even values of 0.2 can be useful.

The correlation between exam results and test score after the first exam sitting for the 728 students for whom data was available was 0.49. This value is highly statistically significant and showed a strong relationship between the test and exam performance. Because it is likely that poorer students are less likely to have had an exam result reported in the sample, the value was considered a possible underestimate the true relationship.

For the post resit group the statistic was calculated on two thirds of the sample, randomly selected and then cross validated on the remaining third. The result for the two thirds of the sample (n=642) was 0.51. It was the same for the remainder of the sample (n=346). For the whole group of 988 the result was the same. These results are highly statistically significant. The 95% confidence interval for the correlation ranges from 0.46 to 0.55. The 99% confidence interval ranges from 0.45 to 0.57. This means that it can be concluded with a high degree of probability that there is a correlation of not less than 0.45 between the test and exam scores.

In order to use the test score to predict outcomes the relationship needs to be expressed as a function. A regression analysis was used to identify the optimal function. The previous report provided an interim analysis. In this phase a cross validation approach is taken. The optimal function is identified using two thirds of the sample and its effectiveness is tested on the remaining third.

Entering just the single variable for test score into the equation results in a multiple R identical to the correlation. The multiple R is an indicator similar to a correlation which can express the relationship between a linear combination of variables and another variable.
The squared multiple R is an estimate of the proportion of variance explained by the predictor variable. In this case the value is 0.24 – 24% of the variation in exam results can be explained by the test. The result is highly statistically significant.

Non-linear regression was explored with quadratic, cubic, logistic and logarithmic equations. However the non-linear solutions were no better than the linear result. The simpler linear model is preferred where alternative models do not improve fit to the data.

Table 3.1 shows the unstandardised coefficients. The unstandardised coefficients represent the multiplier and constant for a linear equation to predict exam results from test score. The standardised coefficient expresses the multiplier in standard deviation units and provides results that are more comparable across variables when more than one predictor is used. The standardised coefficient for test score is just under a half. This means that the exam score is predicted to rise by about half a standard deviation for every standard deviation rise in the test score.

Table 3.1: Regression of Exam results on Test Score

Indicator	First sit Results	Two thirds post resit results	Cross validation sample	Whole sample post resit
N	728	642	346	988
Multiplier	4.578	4.712		4.650
Constant	73.811	72.683		72.972
Multiple R	0.493	0.513	0.507	0.509
Adjusted Multiple R^2	0.242	0.262	0.257	0.258

The multiplier and constant provide a linear equation to convert test scores into predicted exam scores. Table 3.2 shows some example results of applying the prediction equation. The prediction is the best estimate of the future exam score but it will not be exact. There are many factors that affect exam scores apart from ability as measured by the test. The amount of work the student does will have a large impact along with such factors as previous understanding and knowledge, personal situation and interest in the course.

The standard error of estimate (SEE) is the typical difference between the predicted exam score and the actual score achieved. The standard error of estimate in this case is 6.9 for the full sample. This means for any student there is a 67% chance that their actual exam score will lie between one SEE above the predicted score and one SEE below. There is a 96% probability that the actual examination score will be within 2 SEE of the predicted score. Table 3.2 shows the 96% confidence interval for the predicted scores.

If the exam pass score is 60 then a test score of -3 or above would predict this result. While the pass score for each module is 60, students with an average exam score of 60 probably fail since they will have some scores below 60 as well as some above. Table 5.9 shows the range of exam scores associated with each grade. The average score of those graded competent is above 67.

Table 3.2: Predicted exam scores from test scores

Test Score	Predicted Exam	96% confidence interval
-3	59.0	45-73

-2	63.7	50-78
-1	68.3	54-82
0	73.0	59-87
1	77.6	64-92
2	82.3	68-96
3	86.9	73-100

3.2 Predicting Grades

In order to explore the relationship between the test score and the exam grades the test score was split into 5 bands each encompassing approximately 20% of the sample. Table 3.3 shows the comparison between these bands and the grade outcomes. Cells containing more than 25% of a grade band are shaded. This shows that 77% of those who were graded Outstanding were in the top 2 score bands on the test whereas 70% of those who failed, and 64% of those who were referred after resits were in the bottom two score bands.

Considering the breakdown in the other direction, of those scoring in the highest band on the test, 24% are graded outstanding compared to only 9% overall, only 1% fail, 1% are deferred and 5% are referred. In contrast 18% of those in the lowest band on the test fail, 23% are referred and 3% are deferred. Only 28% pass at first sitting.

APPENDIX 5

Table 3.3: Relationship between test scores and grades

	Grade								
Test score range	Outstanding	Very Competent	Competent	Pass year one	Referred	Referred year one	Fail	Deferred	Total
1 Low	6.8%	11.3%	25.0%	15.4%	37.8%	70.0%	43.6%	50.0%	18.9%
2	6.8%	16.0%	29.5%	30.8%	25.9%	10.0%	26.7%	7.1%	19.5%
3	9.4%	23.3%	22.7%		16.3%	10.0%	23.8%	7.1%	21.2%
4	23.1%	25.7%	16.4%	15.4%	11.1%	10.0%	3.0%	14.3%	20.3%
5 High	53.8%	23.6%	6.4%	38.5%	8.9%		3.0%	21.4%	20.1%
Total (100%)	117	661	220	13	135	10	101	14	1257

Table shows Percentage of each grade scoring within score range

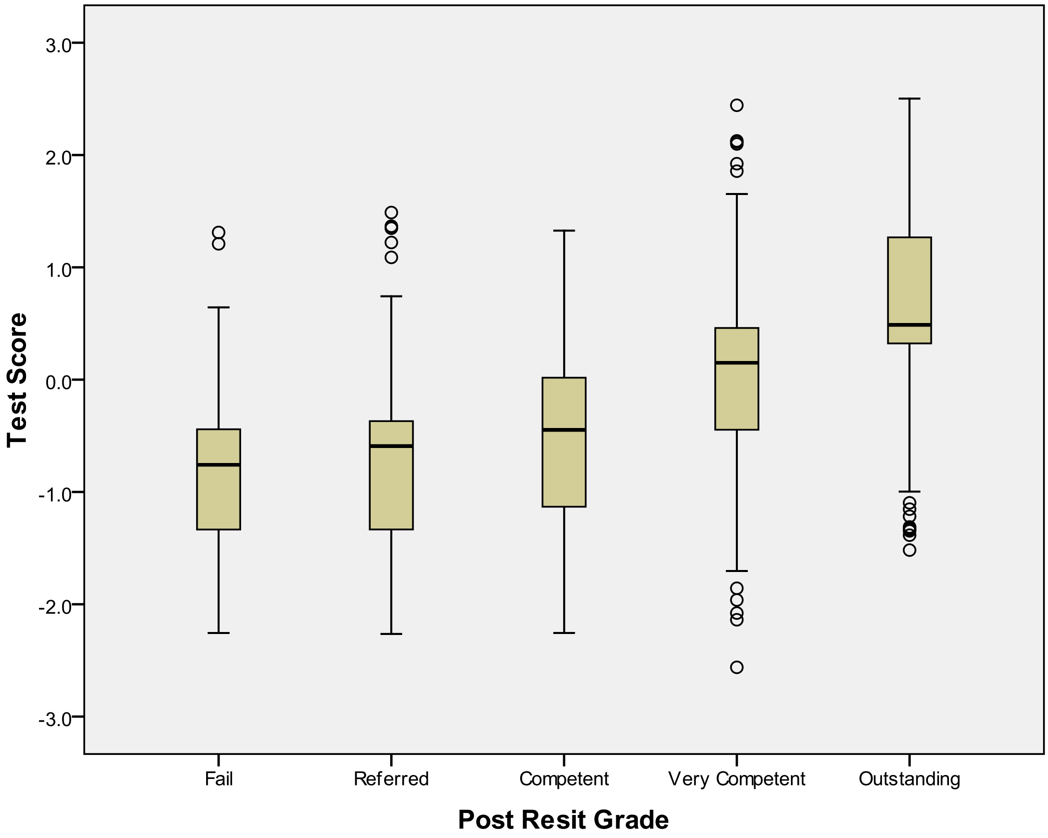
The relationship is not perfect. There is a small percentage of those graded Outstanding or Very Competent that are in the lowest 2 bands of the test. Some of these may be students who did not invest any effort in completing the test and therefore have scores which do not reflect their ability. However the prediction from the test will not be perfect and although these results show there will be a substantial improvement in outcomes if the test were to be used in selection, its use will exclude occasional students who would have done well on the course. Of course this will be true of any selection rule. There will be some people with a third class degree who might have done well on the course. Like those with low test scores, they will be exceptions to the general rule that those with better qualifications tend to do better on the course. Section 5 provides results regarding the effectiveness of applying a cut based on the test scores.

Figure 3.1 is a box plot showing the spread of test scores at each outcome grade. The box in the middle of the column shows the middle 50% of scores. The line through the box is the median (middle score). The whiskers above and below show the range of the remainder of the scores although exceptional outliers are shown as separate circles.

When comparing 2 groups if the two boxes are on about the same level then the score ranges are similar. If the median lines are also on the same level then the groups are very similar. Differences in the whiskers are less important but indicate a larger or smaller range of scores for one group.

While there is a wide spread of test scores for each outcome there is a clear relationship between the test scores and outcome grade.

Figure 3.1: Box plot showing test score ranges for each outcome category



4. Incremental Validity

The previous section shows that the test score is predictive of examination results. However it may be the case that other information could do this equally as well. In the current data set there is information about A level (or equivalent) results and first degree grades and institution of study that could provide similar information. This information is only available for part of the sample. Using this data a two phase multiple regression analysis was performed to look at the incremental prediction of the test after these factors have been taken into account. A similar approach could be undertaken with IELTS scores but these are not available in the current data set.

Table 4.1 shows the means, standard deviations and samples available for each of the variables for this analysis. As discussed earlier the A level points data may have low reliability because many students had only a single A level result recorded which would be insufficient for university entrance without other qualifications. The institution of study for first degree was coded as 1 if it belonged to the Russell Group and 0 otherwise.

Table 4.1: Mean and standard deviation for regression variables

	Mean	**SD**	**N**	**Correlation with Exam result**
A Level Points	245	177	313	0.19**
Degree Class	8.0	1.2	124	0.46*
Attended Russell Group University	0.56	0.50	132	0.31*
Exam result	72.5	8.3	313	n/a
Test Score	-0.05	0.85	313	0.52**

**p<0.01

The most appropriate regression model evaluates the impact of A level points, institution and degree class in predicting results and then the incremental value of the test scores above this. However because only a small part of the sample had data for all the relevant variables additional analyses were run considering institution, degree class and A level points separately for which larger samples were available. The results are provided in table 4.2

Model 1 is the original model with just the test scores used as a predictor and these results repeat those presented earlier. Model 2 looks at the value of the test after A level points have been taken into account. Model 3 considers the test value after degree class is taken into account. Model 4 considers the test value after first degree institution is taken into account. The last model looks at the value of the test after both degree class and A level points are considered.

The standardised regression coefficients (beta) for each variable are provided as these provide an indication of the relative importance of the different factors. In all the models the test score has the highest beta coefficient which is always statistically significant. A level points and Degree Institution both have significant coefficients when alone in a model but their coefficients do not reach significance in the model with all the predictor variables. While the coefficient for A level points is quite small in the combined model, the coefficient for institution retains its size but no longer reaches statistical significance with the smaller sample available for the combined model. Degree class reaches significance both alone and in the combined models. Overall the test score is shown as the strongest single predictor with degree class as the second best predictor of examination results.

The Multiple R and Adjusted R-Square increase quite substantially as variables are added to the equation. The conclusion is that while the test is the strongest single predictor, including educational information together with the test will provide better prediction than the test alone. In particular degree class seems to be the strongest indicator after test score in the full sample. As the multiple R increases the Standard Error of Estimate reduces. This shows how the accuracy of the predicted examination result increases with the inclusion of more variables in the model.

The incremental R-squared for the test scores is large and always statistically significant. This shows that the test improves the prediction of examination scores by a substantial amount even when educational qualifications are taken into account.

While the effectiveness of previous educational achievements as a selection variable may be relevant to the individual courses, it is less relevant here unless the Bar Standards Board wishes to change the requirements relating to these variables. However the comparison does show that the test is a more powerful predictor of course exam grade than any other variable studied here.

Table 4.2: Comparison of Models to predict Exam score

Predictors	**Model 1** **Test Only**	**Model 2** **A level + Test**	**Model 3** **Degree + Test**	**Model 4** **Institution + Test**	**Model 5** **A level, Degree, Institution + Test**
Sample Size	988	313	124	132	95
A Level Points Beta		0.125*			0.066
Degree Class Beta			0.278**		0.243*
Degree Institution				0.156*	0.156
Test Score Beta	0.509**	0.503**	0.419**	0.472**	0.443**
Multiple R	0.509**	0.535**	0.594**	0.545*	0.635**
Adjusted R-squared	0.258**	0.281**	0.342**	0.286*	0.376**
Incremental R-squared for test	0.258**	0.249**	0.144**	0.198**	0.093**
Standard Error of Estimate	6.94	7.05	5.42	5.68	5.19

* p<0.05; **p<0.01

5. Selecting a Cut Score

The results presented in this section, together with those from the previous section can inform cut score decisions. The tables shows the impact of a range of cut scores on the proportion of students passing, failing or being referred. In the current data set 101 students are identified as failing and 135 are referred after resits.

5.1 Reducing the failure rate

The aim of introducing the test is to prevent students that do not have the necessary ability to complete the course effectively from taking the course. Therefore this section explores the impact of a variety of test scores in reducing the failure and referral rate and increasing the pass rate.

Table 5.1 shows the impact of applying different cut scores to the current sample. Part time students have been excluded from this analysis as it is not yet known whether they will pass the whole course. This leaves a total of 1234 students with identified grades. The bottom row shows that by selecting the best 72% of the sample based on their test scores (Cut score -0.5) the failure rate can be reduced by over 40% (from 8.2% down to 4.8%), the referral rate by 35% (from 11% down to 7%) and the pass rate increases by 9% (from 81% to 88%). Alternative results for lower (less selective) cut scores are also shown.

Table 5.1: Impact of Cut Score on course pass rates after first resits

Cut Score	% Students Selected	% Passing Course	% Passing above Competent	% Referred	% Failing
No Cut Score	100%	81%	63%	11%	8.2%
Cut score -1.5	97%	82%	64%	11%	7.5%
Cut score -1.34	90%	84%	66%	9%	7.0%
Cut score -1.25	84%	85%	69%	9%	6.1%
Cut score -1.0	79%	86%	70%	8%	5.6%
Cut score -0.75	77%	87%	71%	8%	5.3%
Cut score -0.5	72%	88%	73%	7%	4.8%

Tables 5.2 to 5.7 provide a more detailed breakdown of the potential impact of various cut scores. The highest cut score (-0.5) shown in table 5.2 results in over a quarter of those currently on the course failing the test. Of those who fail the exam or are referred again after resits less than half would have passed the test with this cut score. Those who received a marginal pass on the course graded Competent had a 60% chance of passing the test. In contrast 90% of those who obtained an Outstanding grade and 83% of those who were graded Very Competent would have passed the test. This cut score therefore results in a substantial reduction in failures and referrals but also more than 15% of students that passed the course with a grade of Very Competent or Outstanding failing to be accepted onto the course. This figure may be somewhat exaggerated due to some students not taking the test as seriously as they would have, had their access to the course depended on it.

Table 5.2: Predicted test success rates at each grade with a cut score of -0.5

Grade	Fail Test	Pass Test	Total (100%)
Outstanding	10%	90%	117
Very Competent	17%	83%	661
Competent	39%	61%	220
Referred	53%	47%	135
Fail	57%	43%	101
All Grades	**28%**	**72%**	**1234**

Table 5.3 shows the impact of lowering the cut score slightly to a value with a pass rate of over 75% overall. This still results in a 50% reduction in the number of failures and over 43% reduction in referrals. The number of false negatives, those who achieve a high grade on the course despite failing the test is a little lower at 14% overall.

Table 5.3: Predicted test success rates at each grade with a cut score of -0.75

Grade	Fail Test	Pass Test	Total (100%)
Outstanding	9%	91%	117
Very Competent	15%	85%	661
Competent	31%	69%	220
Referred	43%	57%	135
Fail	50%	50%	101
All Grades	**23%**	**77%**	**1234**

Table 5.7 shows the impact of the lowest cut score (-1.5). Only 3% of students would have failed this cut score so it has very little impact on the pass rate generally although 11% of people who failed the course would have failed the test.

Tables 5.4 to 5.6 show the impact of intermediate cut scores which would have resulted in 21%, 16% and 10% of the students being unable to take the course. They both reduce the failure and referral rate substantially although not as much as the higher cut scores. The have a lower impact on those with the best course outcomes with 5%, 9% and 13% false negatives for a cut score of -1.34, -1.25 and -1 respectively. These cut scores seem to provide a better balance of reducing failures without excluding people who might have passed the course.

Table 5.4: Predicted test success rates at each grade with a cut score of -1.0

Grade	Fail Test	Pass Test	Total (100%)
Outstanding	8%	92%	117
Very Competent	13%	87%	661
Competent	28%	72%	220
Referred	39%	61%	135
Fail	46%	54%	101

All Grades	**21%**	**79%**	**1234**

Table 5.5: Predicted test success rates at each grade with a cut score of -1.25

Grade	Fail Test	Pass Test	Total (100%)
Outstanding	5%	95%	117
Very Competent	10%	90%	661
Competent	22%	78%	220
Referred	34%	66%	135
Fail	38%	62%	101
All Grades	**16%**	**84%**	**1234**

Table 5.6: Predicted test success rates at each grade with a cut score of -1.34

Grade	Fail Test	Pass Test	Total (100%)
Outstanding	3%	97%	117
Very Competent	5%	95%	661
Competent	11%	89%	220
Referred	23%	77%	135
Fail	23%	77%	101
All Grades	**10%**	**90%**	**1234**

Table 5.7: Predicted test success rates at each grade with a cut score of -1.5

Grade	Fail Test	Pass Test	Total (100%)
Outstanding	1%	99%	117
Very Competent	1%	99%	661
Competent	4%	96%	220
Referred	4%	96%	135
Fail	11%	89%	101
All Grades	**3%**	**97%**	**1234**

As well as reducing failure rates overall, it would be desirable if a larger proportion of students passed at first sit examinations without the need for any resits. Table 5.8 shows the reduction in those not passing at the first sit for various cut scores. The lowest cut score reduces the number of failures and referrals at first sitting by around 15%. The highest of the three scores reduces the failures at first sitting by 40% and the large number of referrals by 30%. There is a bigger impact on those referred who go on to fail. Just over half these students would have passed the test. In contrast over 80% of those who were referred but go on to pass the course well would have passed the test with the highest of these cut scores.

Table 5.8: Percentage Passing Test with Cut between -1 and -1.34

Grade at second sit Grade at first sit	Fail second sit	Referred second sit	Competent second sit	Very Competent second sit	Outstanding second sit	Total
Fail first sit						
Number	29	1	7	5		42
Percentage passing test						
Cut= -1.34	83%	0%	100%	100%		86%
Cut= -1.25	66%	0%	71%	100%		69%
Cut= -1	52%	0%	71%	100%		60%
Referred first sit						
Number	70	131	162	133	1	497
Percentage passing test						
Cut= -1.34	74%	78%	86%	93%	100%	84%
Cut= -1.25	61%	67%	74%	87%	100%	74%
Cut= -1	56%	62%	69%	82%	100%	69%

Figure 5.1 compares the distribution of grades for the unselected and selected group with the 3 cut scores. The blue area represents students who would not have been accepted on the course with any of the cut scores. The green band represents students who would have passed the lower cut score but not the next one. The beige band represents students who would have failed the highest cut score. The figure shows that the use of any of the cut scores has a substantial impact on the failure and referral rate but little impact on the number of Outstanding results.

5.2: Direct Test Score Bands

Another approach to determining a cut score is based on identifying the minimum test score associated with desirable outcome grades. The desirable outcome grade could be set at a marginal pass (Competent) or at a good pass level (Very Competent). While Competent represents an adequate level to pass the course and go on to practice, it is a relatively narrow grade. It could be regarded as more desirable to aim for students to pass at the ‘Very Competent’ level even though some will pass at the lower grade. The following discussion is based on setting the level of the test to be consistent with an aim of students attaining a Very Competent or above outcome.

Figure 5.1: Distribution of grades with 3 cut scores

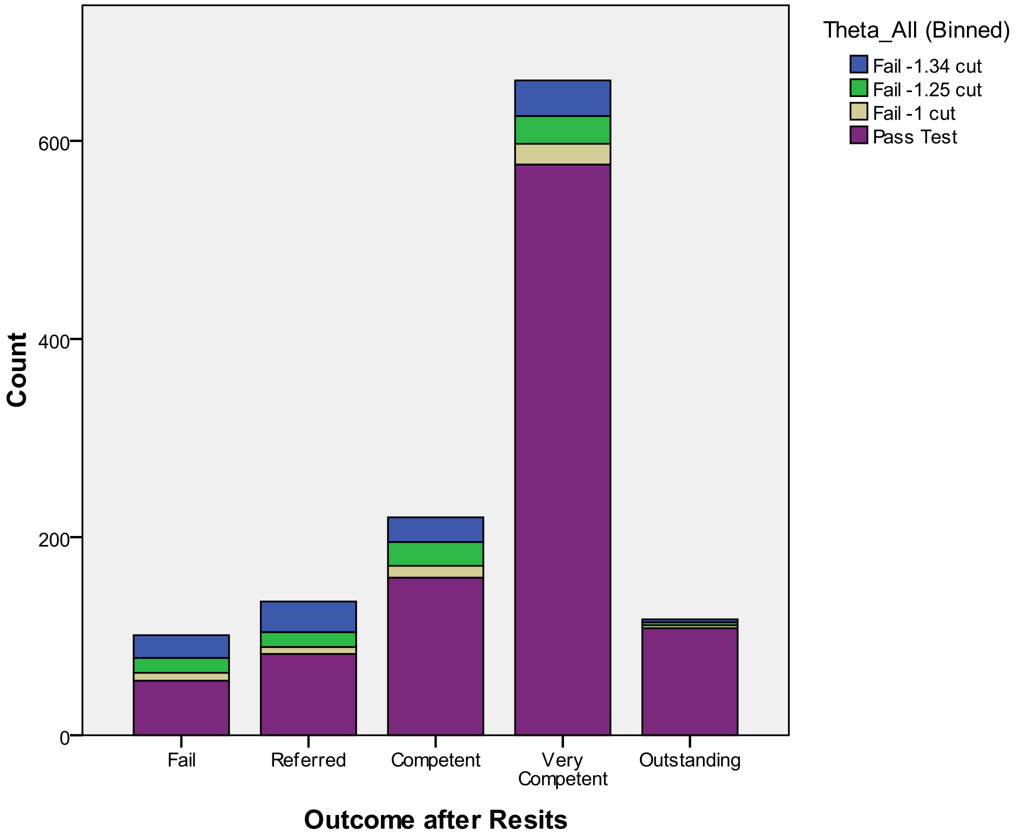


Figure 3.1 shows that there is a very broad range of test scores associated with each grade. Table 5.9 shows the broad range of test scores associated with each grade outcome in figures. There is a large overlap between the categories. This can be seen in the low minimum scores for every category and in the minimal reduction in the standard deviation for categories compared to the full sample. However the large degree of overlap may be partly due to outliers. A few students with extremely high or low scores can increase the range of scores observed in each category even if the majority of scores are more closely clustered. To address this, instead of using the maximum and minimum score the 5th to the 95th percentiles are used as the boundaries for categories. This still reflects the broad range of scores for the category but ignores outliers.

The minimum test score attained by someone achieving a Very Competent outcome is -2.56. This is the minimum test score of the whole sample so is not an effective cut score. Using the 5th percentile for this Very Competent category ignores the lowest scorers on the test who may be outliers. This provides a minimum test score of -1.34.

Table 5.9: Test score by grade outcome

Grade	Fail	Referred	Competent	Very Competent	Out-standing	All
Mean test score	-0.84	-0.61	-0.47	0.03	0.55	-0.15
Standard Deviation	0.68	0.8	0.72	0.8	0.86	0.87
Minimum	-2.26	-2.26	-2.25	-2.56	-1.52	
5th percentile	-2.10	-1.48	-1.46	-1.34	-1.31	
95th percentile	0.38	0.81	0.53	1.32	1.81	
Maximum	1.31	1.49	1.33	2.44	2.50	
N	101	135	220	661	117	1234

5.3: Predicting a test cut score from an exam score

There is a stronger relationship between the test scores and the exam results and the exam results are more closely related to the outcome grades. Figure 5.2 shows the box plot of results for exam scores against outcome grade. The score ranges are more tightly clustered than for the similar chart for the test scores and there is a little less overlap. If a desired minimum exam result can be identified, the prediction equation identified in section 4 can be used to identify the test score which is associated with that examination score.

Figure 5.2: Box plot of exam scores by grade

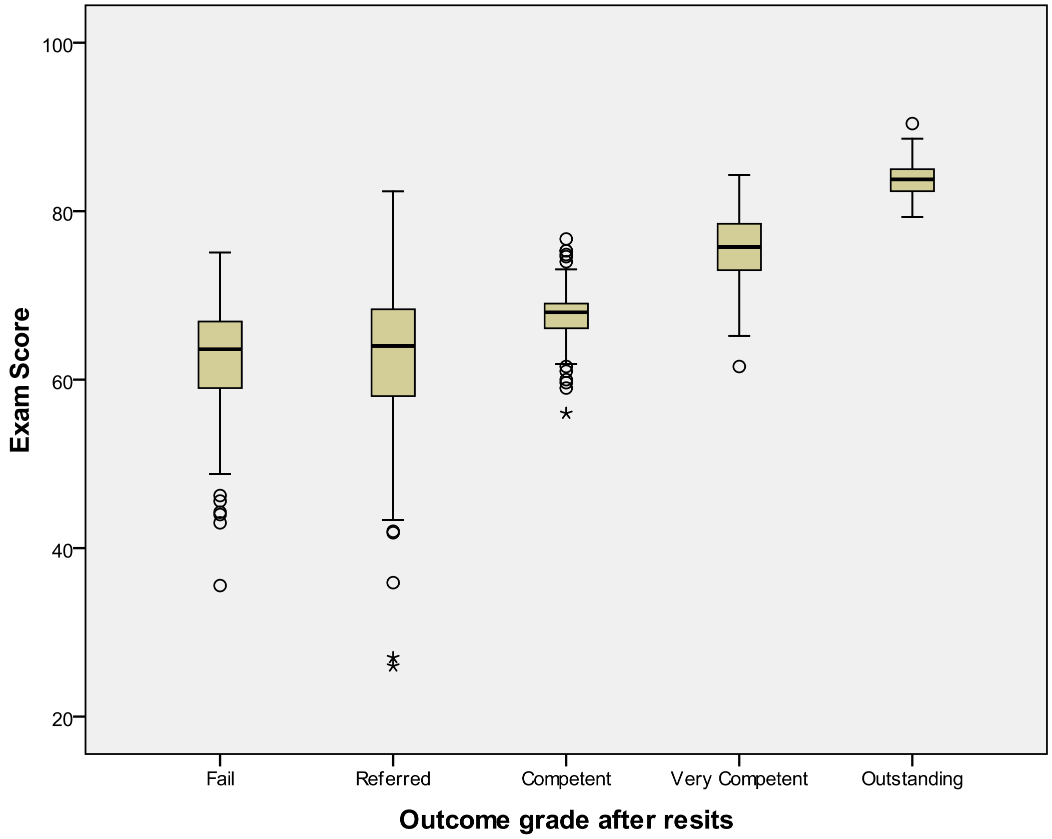


Table 5.10 shows the range of exam scores associated with each course grade. The minimum exam score associated with a grade of Very Competent is 61.6 which is very similar to the 5th percentile score associated with the Competent grade (62.2). The 5th percentile exam grade for Very Competent is 70.5. Having identified these exam scores the next step is to find the test score which is associated with each.

Table 5.10: Exam scores by grade outcome

Grade	Fail	Referred	Competent	Very Competent	Out-standing	All
Mean test score	61.9	62.1	67.7	75.8	83.8	72.1
Standard Deviation	7.2	9.9	3.0	3.5	2.1	8.1
Minimum	35.6	26.0	56.0	61.6	79.3	
5th percentile	45.6	42.3	62.2	70.5	80.6	
95th percentile	71.1	73.8	72.8	81.5	87.3	
Maximum	75.1	82.4	76.7	84.3	90.4	
N	101	104	209	661	88	1035

Table 3.1 provides an equation for predicting exam scores from test scores. This equation can be used to find the test score which predicts the required exam score.

Table 5.11 shows the test scores that best predict these examination outcomes. The test scores which predict the first two exam scores are very low and if used as cut scores would hardly have excluded any of the current sample. The third exam score is predicted by a test score of –0.53 which is at the upper range of the cut scores considered in section 5.1.

Table 5.11: Predicted exam scores from test scores

Test Score	**Predicted Exam Score**
-2.45	61.6
-2.32	62.2
-0.53	70.5

5.4: Cut Score Summary

Section 5.1 shows that success rates and proportions achieving higher grade passes can be substantially increased by the use of the test. It should be remembered that this sample is pre-selected using existing selection processes so the impact is incremental over current methods of selecting candidates. Even a low cut score on the test results in some improvement. However the impact of the test is probabilistic and there will always be some incorrect decisions. Cut scores in the range -1.5 to -0.5 were examined and scores between -1.25 and -1 provided a marked reduction in students who go on to fail the course without creating an enormous barrier for applicants or excluding many students who had good course outcomes.

Prediction from exam scores suggests a cut score of up to -0.53. Direct inspection of the test scores associated with each grade outcome suggests a cut score of -1.34.

In introducing a new test it is desirable to take a conservative approach in setting a cut score until there is clear evidence of its effectiveness. The current results show that the test is an effective predictor of performance for current students, however there may be some differences in the results for applicants when the test is operational. For this reason a cut score toward the lower end of the identified bands is suggested initially. Following the introduction of the test with applicants and validation in operational use, it would be possible to increase the cut score if this was indicated.

Another reason to opt for a cut score towards the lower end of the identified band is that the purpose of the test is not to increase the overall performance of students on the course but to prevent those who are most likely to fail from wasting money and effort on the course with little hope of a useful outcome for themselves.

For these reasons a cut score between -1.34 and -1.25 is suggested. Section 7 looks at the impact of these cut scores on different demographic groups.

6. Group Comparisons

The monitoring report provided a breakdown of test scores by some of the demographic variables before the exam results were available. The analysis showed small differences for most comparisons but larger differences with respect to Ethnic Group. Differences in test scores will lead to adverse impact when the test is used in selection. Such impact can only be justified if it can be shown that the differences are related to course outcomes for that group. The statistical approach to looking at this is a hierarchical regression somewhat similar to the incremental approach used in section 4. If including the group membership variable in the regression equation significantly improves prediction it suggests some bias in the results. Both the group membership variable and its interaction with test score are tested. If there is no improvement in prediction then the score differences on the test are also reflected in the outcome.

Hierarchical analyses were presented for those variables for which the sample was large enough in the previous report. Here the analyses have been repeated where indications of difference were found or the results were inconclusive because of small samples.

6.1 Regression Results –Gender

Although there is small but statistically significant difference in test scores with men scoring marginally higher than women, the regression analysis showed no evidence of any bias in the test in predicting course outcomes.

6.2 Primary Language

A small but statistically significant difference in test score between those who have English as a primary language and those who do not was seen in previous reports, however exam results were only available for 35 people who reported English as a second language and this is insufficient to perform the analysis. Following resits this sample is now 49. This is still insufficient for the hierarchical analysis but Table 6.1 updates the results presented previously for the larger sample.

There remains a larger difference in exam results to that found with the test although the difference has reduced a little. While those who do not have English as a primary language perform less well on the test, their exam scores are even lower relative to the majority group. The test is therefore not overestimating potential performance differences.

The correlation between the test score and exam was 0.50 ($p<0.01$) for the group with English as a primary language and 0.40 ($p<0.05$) for the group with non-primary English. While both these results are statistically significant in their own right they are not significantly different from each other. The test is clearly predictive for both groups although results with a larger sample are desirable.

Table 6.1: Comparison of Test Results by primary language

	English Primary Language	Other Primary Language
Sample Test	867	80
Test Mean	-0.09	-0.33
Test Standard Deviation	0.86	0.79
Standardised Difference	0.28*	
Sample Exam	567	49
Exam Mean	72.9	69.3
Exam Standard Deviation	7.8	6.1
Standardised Difference	0.47**	

* p<0.05; **p<0.01

6.3 Regression Results - Age

The previous report found a moderate difference between some age groups with the older students performing less well on the test. These groups also tend to have lower scores on the course exam. In the regression analysis the model including the age variable did provide a small but significant improvement in prediction. The analysis was repeated with the larger sample following resits and the outcome was similar. There is a statistically significant but small direct impact of age. Younger students are predicted higher exam scores for the same test score. The size of the effect was reduced in this analysis including the resit data. There is no significant interaction effect. Although the effect does reach statistical significance it is actually small and will have little or no meaningful impact on decision making.

Table 6.2 Hierarchical Regression - Age

Predictors	Model 1	Model 2	Model 3
Sample Size	642	642	642
Test Score Beta	0.497**	0.360**	0.49**
Age Beta		-0.09**	-0.11**
Age*Test Beta		0.144	

Multiple R	0.497**	0.512**	0.508**
Adjusted R-squared	0.245**	0.258**	0.256**
Incremental R-squared for Age and Age*Test	n/a	0.015**	0.003

* p<0.05; **p<0.01

6.4 Regression Results – Ethnic Group

The previous analysis found substantial differences in both test and exam scores for those from different ethnic backgrounds. Only the Asian group was large enough to consider separately and the regression analysis showed a significant effect. Asians tended to score lower than White students on both the test and the exams. The hierarchical analysis showed significant improvement in prediction with both the main and interaction effects reaching statistical significance, although paradoxically the test tended to favour the lower scoring Asian group. There were similar findings for a combined ethnic minority group.

There has been some increase in sample size with the inclusion of resit results therefore the analyses have been repeated here. Exam results were available for 348 students who described themselves as white and 270 from other ethnic groups including 173 from various Asian groups. Again only the Asian group is large enough to consider separately

Table 6.3 shows the results for the White Asian Comparison. There is a significant improvement in prediction by including the main effect of the ethnicity variable but with this larger sample the interaction term does not reach significance. The main effect for test score remains the strongest predictor in the model. This means that separate regression equations for White and Asian students provide better overall prediction than the combined regression line. However the test does predict exam performance well for both groups.

Table 6.3 Hierarchical Regression – White versus Asian

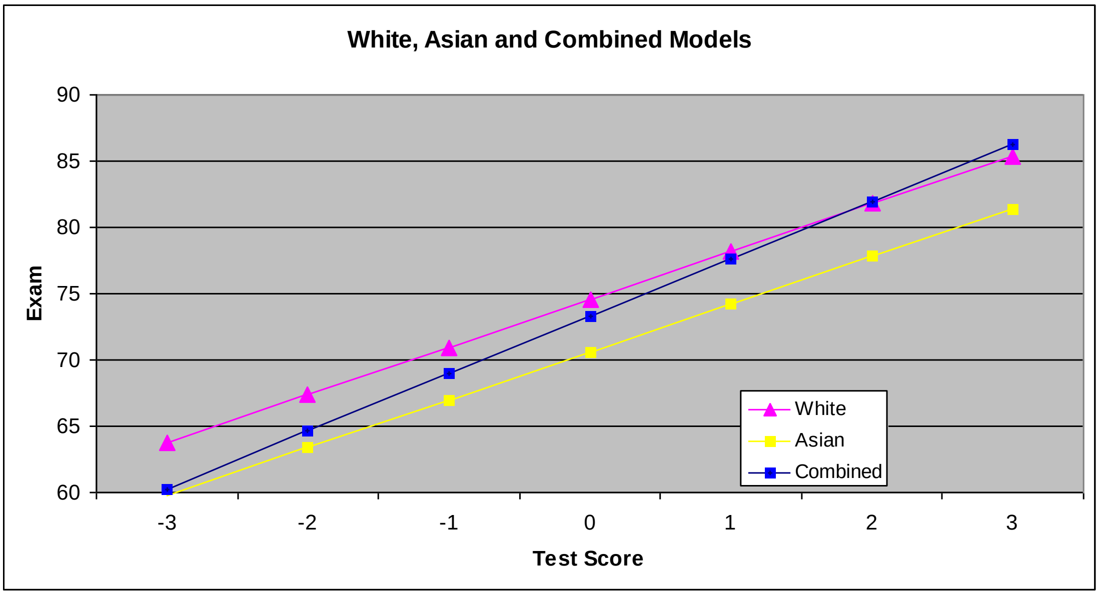
Predictors	**Model 1**	**Model 2**	**Model 3**	**Model 2**
Sample Size	521	521	521	521
Test Score Beta	0.493**	0.277*	0.411**	0.277*
White/Asian Beta		-0.230**	-0.247**	-0.230**
White/Asian *Test Beta		0.147		0.147

Multiple R	0.493**	0.547**	0.545**	0.547**
Adjusted R-squared	0.241**	0.295**	0.294**	0.295**
Incremental R-squared for White/Asian and White/Asian *Test	n/a	0.056**	0.054**	0.002 Increment of Model 2 over Model 3

* p<0.05; **p<0.01

Figure 6.1 shows plots of the different prediction lines. Although the Asian group have a lower average score than the white group, at the same test score, an Asian student would be predicted a lower exam score than a White student with the same test score. This means that using the combined prediction line, shown in blue in the figure, would tend to over-predict the performance of Asian students, particularly in the higher score ranges and under-predict that of white students, particularly in the lower score ranges. This suggests that if there is bias in the test it is against the higher scoring White group rather than the lower scoring Asian group.

Figure 6.1: Comparison of regression equations for separate and combined models



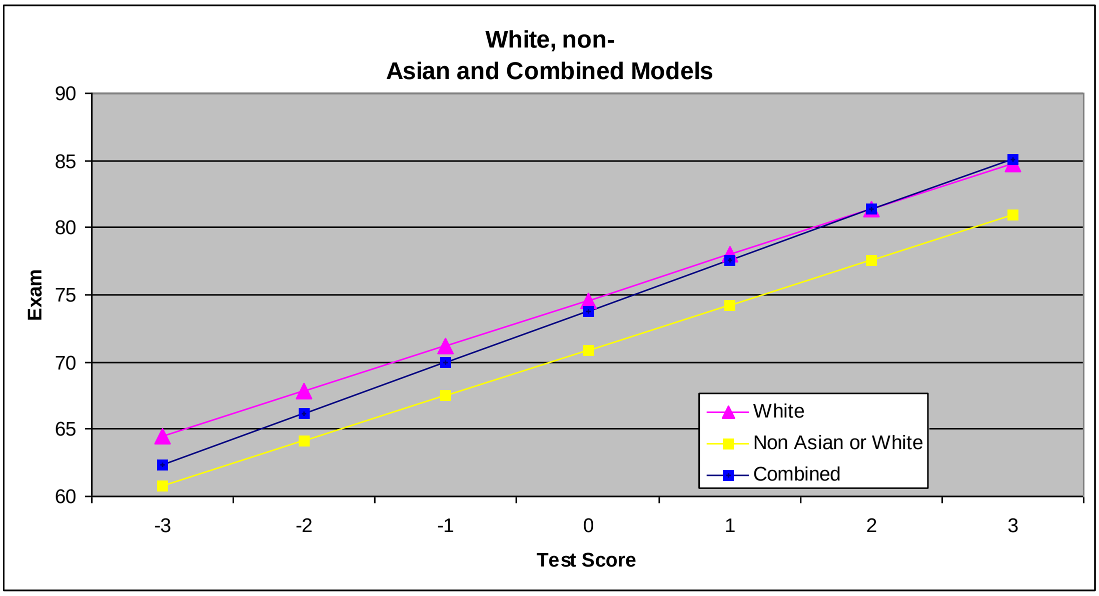
A similar analysis was performed for a comparison between the White group and the remaining (non Asian) ethnic groups. Table 6.5 shows the regression results. The pattern is very similar to that for the White-Asian Comparison. There is a main effect of ethnic group but no interaction effect and the combined prediction equation over-predicts the exam performance of the ethnic minority group. Figure 6.2 compares the regression equations.

Table 6.4 Hierarchical Regression – White versus non Asian Other

Predictors	**Model 1**	**Model 2**	**Model 3**
Sample Size	444	444	444
Test Score Beta	0.452**	0.400**	0.378**
White/Other Beta		-0.212**	-0.300**
White/Other *Test Beta		0.002	
Multiple R	0.452**	0.497**	0.497**
Adjusted R-squared	0.203**	0.242**	0.244**
Incremental R-squared for White/Other and White/Other *Test	n/a	0.043**	0.043**

* p<0.05; **p<0.01

Figure 6.2: Comparison of regression equations for separate and combined models



6.5 Disability

The previous report showed no statistically significant difference in test score between those who have a disability and those who do not but a significant difference in exam scores with non disabled students tending to score around a third of a standard deviation higher in exams. However exam results were only available for 33 people with a variety of different disabilities. With the resit information this increases to 48 which is still insufficient for a full analysis but table 6.5 has been updated.

Following resits the difference in exam scores between disabled and non-disabled students disappeared and there is now no significant difference between the groups on either the test or the exam.

The correlation between the test score and exam was 0.52 for the non disabled group and 0.50 for the disabled group. Both these results reach statistical significance at the 0.01 level. This suggests that the test is predictive for both groups although results with a larger sample are desirable. In particular it would be informative to consider different disabilities separately. The current group includes 20 people with dyslexia with the remainder describing a range of different disabilities.

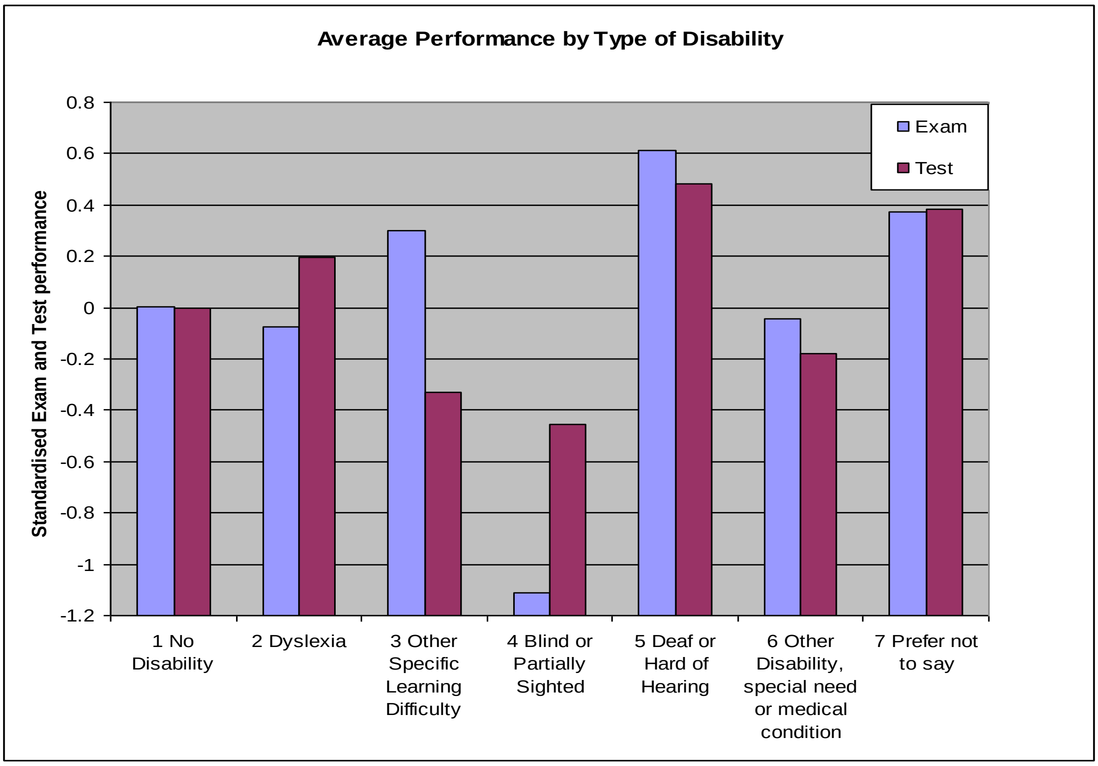
Table 6.5: Comparison of Test Results by disability

	No Disability	**Disability**
Sample Test	783	75
Test Mean	-0.14	-0.19
Test Standard Deviation	0.87	0.91
Standardised Difference	0.06	
Sample Exam	508	48
Exam Mean	72.5	71.5
Exam Standard Deviation	7.5	8.4
Standardised Difference	0.13	

* p<0.05; **p<0.01

Figure 6.3 shows the mean scores on the Exam and Test broken down by disability. There are some noticeable differences between difference disabilities with, for example, the Deaf and Hard of Hearing group scoring much higher on both measures than the Blind or Partially sighted group. However these groups are very small so it is not possible to draw strong conclusions from these results.

Figure 6.3 Standardised test and exam scores by disability



As the number of disabled students will always be quite small, it is recommended that more qualitative techniques are used to ensure that accommodations for disabled applicants meet their needs and that they are not disadvantaged by the nature of the test or how it is administered.

6.6 Regression Results – University type

Previous analyses have found large difference between the test scores for Russell Group graduates over other university graduates. This is likely to be related to the greater selectivity of the Russell group universities in choosing students initially. In this data set those who had attended Russell Group institutions had an average of 90 more A-level points than those who had attended other universities.

The regression analysis has been repeated as the sample is now 30% larger. As before the model including the university type variables did not result in any significant improvement in prediction. The implication is that the higher average test scores of Russell Group applicants are as a rule reflected in their performance on the course.

Table 6.6: Hierarchical Regression – Type of University

Predictors	**Model 1**	**Model 2**
Sample Size	132	132
Test Score **Beta**	0.524**	0.549**
University Type **Beta**		0.142
University Type ***Test Beta**		-0.091
Multiple R	0.524**	0.548**
Adjusted R-squared	0.270**	0.283**
Incremental R-squared for University Type and University Type*Test	n/a	0.025

* p<0.05; **p<0.01

7. Example Impact of Cut Score

Where there are score differences between groups this will be reflected in the results of applying a cut score. A smaller proportion of the lower performing group will pass any particular cut. The following tables show the impact of using a cut score of -1.34, -1.25 and -1 on the test on different groups.

The tables also show the relative selection ratio for the groups compared. This is the ratio of the pass rate for the group with the lower success rate to that with the higher pass rate. Where this value is below 0.8 the selection fails the 'four fifths rule' and is considered to have significant adverse impact. The implication if this is the case in a real selection process is that people from the lower scoring group have less than 80% the chance of people from the higher scoring group of being offered a place to study.

The comparison by gender shows only minor difference in pass rates for these cuts scores despite a small score difference. All the selection ratios are well above 0.9. The same is true for the Primary Language, and Age comparisons despite small differences in test scores for these groups. The pass rate for the disabled group is marginally higher than for the non-disabled group. While this difference is not meaningful it does show that overall disabled people do not seem to be disadvantaged by the test.

Even for the Ethnic Group comparisons where the largest differences in test scores were seen, all of the selection ratio comparisons are in the target region of 0.8 or above. However the results for the highest cut score (-1) for the Asian group are on the border of the target zone.

The pass rates by university type show a similar pattern. At the highest cut score the relative selection ratio is just above 0.8.

The pass rate is consistently higher for those who go on to pass the exam than for those who are referred for resits or who fail except one or two cases where the pass rate is based on a very small number of students. This means that using the test will result in an increased pass rate as discussed in previous sections.

Minimising the degree of adverse impact should be one of the criteria in choosing a cut score. All three cut scores explored here show none or minor levels of adverse impact which is within acceptable parameters with respect to all the facets studied.

Table 7.1: Pass Rate by Gender

% passing test	Test Pass Rate Cut=-1.34		Test Pass Rate Cut=-1.25		Test Pass Rate Cut=-1	
	Male	Female	Male	Female	Male	Female
Sample	366	461	366	461	366	461
Course						

Outcome						
Pass Exam	96%	96%	91%	87%	89%	83%
Referred	75%	81%	64%	72%	57%	65%
Fail Exam	78%	85%	63%	68%	59%	55%
All	92%	92%	86%	84%	83%	79%
Relative Selection ratio	1		0.98		0.95	

Table 7.2: Pass Rate by Primary Language

	Test Pass Rate Cut=-1.34		Test Pass Rate Cut=-1.25		Test Pass Rate Cut=-1	
	English	Other	English	Other	English	Other
Sample	726	62	726	62	726	62
Course Outcome						
Pass Exam	95%	90%	90%	85%	87%	79%
Referred	78%	83%	69%	83%	65%	50%
Fail Exam	84%	75%	63%	75%	54%	63%
All	92.7%	87.1%	85.5%	83.9%	82%	74.2%
Relative Selection ratio	0.94		0.98		0.91	

Table 7.3: Pass Rate by Age

	Test Pass Rate Cut=-1.34		Test Pass Rate Cut=-1.25		Test Pass Rate Cut=-1	
	Under 30 years	30 and over	Under 30 years	30 and over	Under 30 years	30 and over
Sample	719	101	719	101	719	101
Course Outcome						
Pass Exam	95%	95%	89%	92%	86%	87%
Referred	82%	65%	73%	47%	68%	35%
Fail Exam	85%	70%	66%	60%	58%	50%
All	92.9%	87.1%	85.5%	81.2%	81.6%	74.3%
Relative Selection ratio	0.94		0.95		0.91	

APPENDIX 5

Table 7.4: Pass Rate by Ethnic Group

	Test Pass Rate Cut=-1.34			Test Pass Rate Cut=-1.25			Test Pass Rate Cut=-1		
	White	Asian	Other	White	Asian	Other	White	Asian	Other
Sample	437	233	123	437	233	123	437	233	123
Course Outcome									
Pass Exam	96%	91%	97%	92%	85%	88%	90%	80%	79%
Referred	88%	67%	83%	76%	60%	67%	72%	52%	67%
Fail Exam	100%	77%	74%	77%	62%	58%	77%	42%	53%
All	95.9%	85%	91.1%	90.2%	77.7%	80.5%	88.3%	71%	73.2%
Relative Selection ratio compared to White		.89	.95		.86	.89		.80	.83

Table 7.5: Pass rate by Disability

	Test Pass Rate Cut=-1.34		Test Pass Rate Cut=-1.25		Test Pass Rate Cut=-1	
	None	Disabled	None	Disabled	None	Disabled
Sample	651	55	651	55	651	55
Course Outcome						
Pass Exam	94%	98%	89%	88%	85%	88%
Referred	78%	71%	66%	71%	60%	71%
Fail Exam	82%	86%	62%	86%	51%	86%
All	91.6%	92.7%	83.9%	85.5%	79.4%	85.5%
Relative Selection ratio	0.99		0.98		0.93	

Note that the pass rate for the disabled group is marginally higher than for the non disabled group.

Table 7.6: Pass Rate by University Type

	Test Pass Rate Cut=-1.34		Test Pass Rate Cut=-1.25		Test Pass Rate Cut=-1	
	Russell	Other	Russell	Other	Russell	Other
Sample	81	71	81	71	81	71
Course Outcome						
Pass Exam	95%	91%	92%	80%	90%	76%
Referred	100%	80%	100%	60%	100%	60%
Fail Exam	75%	100%	50%	82%	50%	73%
All	93.8%	91.5%	90.1%	78.9%	88.9%	74.6%
Relative Selection ratio	0.98		0.88		0.84	

8. Psychometric Properties

The work needing to be done by Pearson in creating a bank of questions from which to generate multiple forms of the test is almost complete. I have reviewed the results they obtained to verify that the tests will have good psychometric properties.

The dimensionality of the items sets has been explored and a unidimensional three parameter logistic model with fixed guessing has been fitted to the data. There is currently a bank of some 370 questions available for use. Several hundred new questions are being trialled to expand the question bank further.

Pearson is currently using 335 questions from the bank to randomly generate 40 question tests for operational use. These tests are working well with typical internal consistency reliabilities above 0.80. Alternate form reliabilities ranged between 0.76 and 0.88. There are sufficient constraints on the way tests are generated to ensure they are broadly equivalent. The IRT based scoring process can compensate for any small differences in difficulty between different tests.

I would recommend that the BPTC aptitude test be created in a similar manner but using a slightly longer test. A test with 50 questions would show consistent reliability of 0.80 and above. I would also recommend that for the first year the tests use the currently available question pool with 370 items. With a bank of this size the average overlap between any two randomly selected test forms would be only 1 or 2 questions. This makes it difficult for test takers to pass on useful information to each other. It would require an organised consortium of test takers to gather a useful amount of information.

I would recommend that new questions continue to be developed so that they can be added to the bank and overused questions can be retired. In the first instance the new questions already under development could be used to refresh the question bank after the first year of use. On an ongoing basis it may be desirable to require candidates to complete a slightly longer test where some of the questions are new trial questions under development. A 60 item test would allow each candidate to complete 10 developmental questions. Even with the addition of these extra questions, the test should not take longer than 40 minutes to complete for most candidates.

The current scaling of test scores while entirely adequate for the purposes of this report needs to be converted to an easier to understand scale, both for course providers and for the purposes of feedback to candidates. I would suggest a T-score scale would be appropriate. This scale has a mean of 50 and a standard deviation of 10. Scores would typically range from 25 to 75. A cut score of -1 would equate to a T score of around 40.

Reporting procedures for scores will need to be developed. These will depend on how the test scores will be used. It is possible to report to candidates only whether they have passed or failed the test and are allowed to proceed with their application for a course place. However the scores have much additional value for course providers in making selection decisions and could be shared with them with permission from candidates. Supporting documentation will need to be developed with explanations of score meanings for all those that will receive score reports.

Pearson should be asked to provide a full practice test version which candidates could access via the internet in advance of their formal test to familiarise themselves with the test and question types. The provision of practice versions helps to alleviate any differential benefit of participating in coaching or other activities.

9. Summary and Conclusions

Over a series of three reports the results of the trial and validation exercise for the Pearson test have been explored.

The most important findings are with regard to the ability of the test scores to predict course performance and outcomes. Reports 2 and 3 focused on this issue with the results of report 3 having higher validity since they are based on the larger, broader data set of student performance available following the late summer resits.

Two indicators of course performance were used. The first is the outcome grade and the second is average final exam score. Both showed strong relationships with test scores.

The correlation between test scores and course examination results was 0.51 across the whole sample. This is a very strong result and highly statistically significant. A much smaller correlation would have supported the use of the test in selection. This strong relationship was also found when looking at course outcomes. 54% of those attaining an ‘outstanding’ grade were in the top fifth of test scores whereas 44% of those who failed the course scored in the bottom fifth.

Chapter 4 of this report explored whether the test was a better predictor of course outcomes and performance than other potential indicators such as educational attainments. The findings were that the test was the best single predictor of course outcomes but that other variables can improve prediction when combined with the test score if desired.

The question of selecting a cut score is discussed in section 5. A number of approaches and options are evaluated and a range between -1.0 and -1.34 is recommended.

In evaluating any test to be used in selection the question of fairness is paramount. The monitoring report compared the performance on the test of different subgroups to identify any signs of potential bias. The first three columns of Table 9.1 summarise the results for the different demographic groups. Only the ethnic comparisons found medium or large differences in scores on the test, but there were other smaller differences. Whether these differences are critical depends on whether they are artefacts of the test or whether they relate to overall differences in performance on the course. This was investigated through a hierarchical regression process described in section 6 of this and the initial validation report. The main findings have been entered in the last column of table 9.1. While there are some significant differences there is no evidence that the test would show bias against any of the lower scoring groups. The finding from the initial validation report that the test was more predictive for the Asian group was not replicated with the larger sample here.

Table 9.1 Summary of results regarding demographic groups

Demographic Group	Statistically Significant performance differences on test	Size of Difference	Predictive Difference
Gender	Yes	Small	No
Primary Language	Yes	Small	No but sample small
Age	No	Small-moderate. Younger students score higher.	Small statistically significant difference. Any bias favours older lower scoring group.
Ethnic Group White-Asian	Yes	Large	Significant improvement in prediction. Any bias favours lower scoring Asian group.
Ethnic Group White-Other	Yes	Large	Significant improvement in prediction. Any bias favours lower scoring Non white group.
Disability	No	No overall difference may mask issues for specific groups	Good prediction for both groups
Russell Institution	Yes	Moderate/Large	No difference

Table 9.2 summarises the comparisons based on educational achievements. Rather than being of concern, difference here support the general validity of the measure because they indicated that it is related to educational performance more generally.

Table 9.2 Summary of results regarding educational attainment

Educational Attainments	Statistically Significant performance differences on test	Size of Difference
A level results	Yes	Large
Degree Class	Yes	Large

This report also contains details of modelling potential cut scores and showed that using the test score to select candidates could reduce the proportion of students who fail or are referred. The test can also improve the proportion of students achieving the highest grade and passing the course at the first sitting.

No substantial adverse impact was found using a range of low to moderate cut scores for gender, primary language, age, disability or whether the first degree was from a Russell Group institution. The largest impact was for ethnic group with students from minority groups being somewhat less successful than the white group overall. However even these findings were within the generally accepted range for a cut score.

The final section relates to the psychometric properties of the test.

The monitoring report found a good distribution of scores from the test with the Bar students performing at a similar level to other graduates who had taken the test. Further analysis shows the proposed test has a reliability of at least 0.8. Some development recommendations for the test were made.

Comments on Dewberry's Report: Implications for introduction of the Aptitude Test for the Bar Professional Training Course

26th September

The Evaluation of the Aptitude Test for the Bar Professional Training Course (Phase 2)

Introduction

Dr Chris Dewberry of Birkbeck College, University of London was commissioned to write a report providing an overview of the use of aptitude tests in selection and criteria for evaluating the effectiveness of such a test.

The main part of the report provides a good background to the use of aptitude tests in selection. It covers the history and use of aptitude tests of various kinds, the technical qualities of test and provides case studies of tests currently in use in educational and professional selection contexts. It ends with a presentation of criteria for evaluating Aptitude tests in the legal profession.

This document reviews the plan for implementation of an aptitude test for entry onto the Bar Professional Training Course against these criteria.

Recommended Criteria for Evaluating Ability or Aptitude Tests in the Legal Profession

1. The purpose of the test should be clarified.

The purpose of the BPTC test has been clearly specified. Its aim is to reduce the probability of applicants who do not have an appropriate level of ability to have a fair chance of passing the course from being accepted onto the course. It is therefore aimed at identifying people who fall below a minimum level of cognitive ability. This is somewhat different to many of the tests and contexts discussed in the report where the aim is to identify the best possible performers for a job or course.

The validation of the test should particularly focus, therefore, on its capacity to identify potential failures on the course rather than those likely to achieve an outstanding grade.

2. Establishing content validity

In 2009 a job analysis was undertaken to identify the cognitive requirements of students on the course. This included a review of course requirements and documentation, interviews with teachers and students on the course, and interviews with pupil masters. A number of techniques were used in the interviews. A broad range of evidence was collected to ensure that all facets of the requirements would be considered.

Rather than focusing on characteristics required to perform at a high standard, the job analysis concentrated on characteristics that differentiated students with minimum competence to pass from those at a lower level. It identified a number of cognitive factors that were important for avoiding failure. It is however likely that there will be a large overlap with factors related to excellence. The factors are listed below.

- A logical and structured approach to information
- Accurate analysis of complex information from multiple sources
- Application of rational thought and good judgement under pressure
- Appreciation of nuances in an argument
- Understanding of legal principles
- Drawing sound conclusions from legal and other information
- Correctly identifying the strength and relevance of evidence and arguments
- The application of judgement in addition to more logical analysis

The test was chosen because it provided a good match to several of these factors.

3. Consideration of a range of different techniques

Given the requirement to identify applicants with insufficient cognitive ability to complete the course it is clear that a cognitive measure is required in this case.

While measures of personal style or judgement may well be useful predictors of performance in a legal role they are less appropriate in the current context. It is likely that some personality factors will be related to successful course completion. For example the personality factor of conscientiousness has been found to be predictive of educational performance. However personality factors relate to preferences and style of behaviour and not capacity. Someone who is low on the personality factor of Conscientiousness will not have a natural preference for working hard in a structured manner, but this does not mean they might not do so if required. In contrast, someone who does not have the cognitive capacity to cope with the course content is unlikely to be able to overcome this, even with great effort.

4. Evidence of construct validity

There is no strong need for construct validity evidence in the current context since content validity is good and the use of the test will rest on strong empirical criterion related validity.

That said, the construct validity of the test has been explored by the test publishers and it shows good correlations with other measures of verbal reasoning skills. They have also shown that there are three correlated factors underlying test performance, all of which are relevant to the skills identified as important for passing the course.

5. Reliability

The internal and alternate form reliability will be evaluated as part of the psychometric evaluation being undertaken currently. Alternate form reliability is more relevant in the current context than test-retest reliability over time. Both provide evidence of the stability of test scores over test occasions and since there will be many different forms of the test in use it is particularly important to show that the scores from different forms are equivalent.

6. Criterion for criterion related validity

Since the aim of the test is predict those with potential to pass the course there is a clear criterion – whether students do or do not pass the course. In addition the examination result can also be used as this is a more nuanced indicator of performance on the course differentiating those who just pass the course from those who have very high scores. It also allows the use of standard correlation metrics to evaluate validity

7. Criterion related validity

An initial study of the criterion related validity found a correlation of 0.59 (uncorrected) between test scores and examination results with a sample of 124 students. This is a very compelling result and shows a stronger relationship than typically found. For example Dewberry quotes Stilewell, Delassandro and Reese (2009) who found median correlations of just over 0.3 for the US LSAT test and first year university results.

The criterion related validity of the test is currently being evaluated on a larger sample. Initial results have found a correlation of 0.49 between test scores and examination results for a sample of 728. This is a more than adequate sample in terms of size although does not represent poorer students well as many of these have been referred to re sit and final examination results are not yet available. A further analysis will be undertaken later in the year when re sit results are available. In addition the prediction of failures will be investigated further with this information.

Because this latest study is based on a large sample there is a relatively narrow confidence interval. The 95% confidence interval runs from 0.43 to 0.54. This means that there is a less than a 2.5% chance that the actual relationship between test scores and examination results is less than 0.43.

Both these studies are predictive since the students completed the test some months before they sat their finals. However the students had started the course and there was no selection decision hanging on the test score attained. Once the test is in use it will be possible to undertake a full predictive study with applicants taking the test live as part of the requirements for a place on the course, before they start.

8. Incremental Validity

In one sense the design of the validation studies only shows incremental validity since they are performed on a pre-selected group of students. Information from undergraduate Degree and other educational qualifications is already taken into account. Applicants who did not meet the existing criteria relating to educational qualifications would not be on the course.

A statistical analysis of incremental validity will be performed as part of the current evaluation study. Dewberry (2001) found that both Degree class and university attended significantly predicted course outcome. Initial results show that the test has incremental validity beyond educational qualifications (Degree class and A Level points) and university attended. In fact the test is the strongest single predictor and provides significant incremental validity after both educational qualifications and university attended have been taken into account. The best prediction (highest validity) is achieved by taking all these factors into account together.

9. Subgroup Differences

Subgroup differences with respect to gender, primary language, age, ethnic group and disability have been examined. Data was not available for social class.

No significant differences were found for age, or disability. Statistically significant but small differences were found for gender and primary language. Larger differences were found for ethnic group with white students scoring significantly higher than those from other groups. Similar differences were found for Degree class.

Dewberry focuses on the use of differential item functioning (dif) in addressing group differences. While dif analysis is a recommended part of test development and can flag questions that are potentially biased it rarely accounts for more than a minute part of observed differences in test scores in well developed tests. A more important approach is to analyse differential validity – the extent to which test score differences are reflected in outcome differences, or that predictive validity exists for all subgroups.

Differential validity analyses have been performed for all subgroups where the sample size was sufficient for the currently available data. No significant differences were found for gender.
Significant but small differences were found for age. Larger differences were found for ethnic group with the test being a better predictor of performance for the lower performing ethnic groups. The result suggested that any bias in the test was in favour of ethnic minority groups. Such results are sometimes explained through bias in the criterion variable; however Dewberry (2001) did not find evidence of this for the BVC.

10. Practice Opportunities

Example questions for the test are already openly available. It would be advisable for a full length example test to be provided for candidates to practice before the test goes live. Pearson should be asked to provide content for this as part of the contract. Links to the practice test should be provided from the BSB webpages.

11. Selection Decision Rules

The intention is to provide a minimum passing score on the test for all candidates. This will be based on the validation evidence currently being collected and will therefore be empirically supported. It will be based on identifying candidates who have the minimum ability required to pass the course. Course providers will be free to make further inferences from the test score to support selection decisions if they wish. Rules already exist for using information such as Degree class. While it is possible to provide an optimised prediction score combining information from various sources this is not the responsibility of the Bar Standards Board.

12. Regular Reporting

It would be advisable to institute a regular reporting schedule to review the effectiveness of the test. In the initial period of use of the test reporting should be undertaken annually and include further validation exercises based on operational test use. Over time the frequency of reporting might be reduced, but regular confirmation of the quality, impact and effectiveness of the test should be undertaken.

Conclusions

The requirements set out in the report have either already been addressed in the programme of work or are due to be addressed before the test goes live. Issues still to be addressed include setting the cut score for the test, defining the ongoing reporting procedures and specifying how practice opportunities will be provided. Existing research already provides strong support for the validity of the test but additional evidence is to be collected and analysed which will provide further support for the use of the test.

References

Dewberry, C. (2001) Performance disparities between whites and ethnic minorities: Real differences or assessment bias? *Journal of Occupational and Organisational Psychology,* **74**, 659-673.

Stilwell, L. A., Dalessandro, S. P. & Reese, L. M. (2009) Predictive Validity of the LSAT: A National Summary of the 2007 and 2008 LSAT Correlation Studies. *Law School Admission Council LSAT Technical Report 09-03.*

EQUALITY IMPACT ASSESSMENT

Date of Screening	First drafted 2009, this version 11 June 2010 and updated following receipt of the report on the final pilot November 2011 (which see).
Assessor Name & Job Title	Dr Valerie Shrimplin Head of Education, BSB
Policy/Function to be Assessed	The Bar Course Aptitude Test (‘BCAT’) The proposed change in the entry requirements for the Bar Professional Training Course to require ***all*** students to demonstrate that they have attained the minimum specified threshold on the BCAT. The Test will be available in over 150 centres in the UK and in 165 countries worldwide, at a cost of approximately £60.
Aim/Purpose of Policy	The aim of this proposal is to ensure that only those with adequate critical reasoning, language and other skills can undertake the Bar Professional Training Course (BPTC). This will enable the BSB to fulfil its regulatory role for setting and monitoring standards in relation to entry to the Bar and education and training for Barristers. **Existing Entry Requirements** The new Bar Professional Training Course was put in place from Academic Year 2010-11. As a result of the Wood Review of the BVC (2007-08), standards on the course were raised, including at entry, on exit, and during the course itself. The previous policy regarding entry requirements specified academic and degree qualifications, in addition to Inn membership, as specified in the Bar Training Regulations (BTR18 - BTR26). The present entry requirements have proved unworkable, for a variety of reasons. **Rationale for propose new policy** Practice at the Bar demands a high level of ability in critical reasoning, including the written and oral use of the English language. Experience has shown that not all those who meet the current entry requirements are suitable to undertake the bar Course, as demonstrated by the high first time fail rates, which have increased with the raised standards of the new BPTC. The Test is not designed specifically to cap numbers but to ensure the legitimate aim that only suitable

	candidates with a reasonable prospect of passing undertake the course. This is important not only in terms of protecting weak students from the high cost of failure but also because the presence of weak students impacts on the learning experience of others on this highly interactive course. Consequently, students enrolling on the BPTC must have good analytical and critical reasoning skills and also be fluent in English, conversant with the rules of grammar and able to express themselves clearly and accurately both when they speak and when they write. The existing entry requirements have not been able to achieve this. In addition, the imposition of a language test either selectively or universally has proved problematic since it could be viewed as either discriminatory (if required of certain candidates) or disproportionate (if required of all applicants. In addition, language testing does not cover the important analytical and critical reasoning skills that are also necessary. This is why the recommendation of the Working Group was to develop an Aptitude Test that would be the same for all applicants. **Proposed Changes** It is therefore proposed that, to avoid discrimination and aim to be fair to everyone, all applicants must take the Bar Course Aptitude Test (BCAT) and demonstrate that they have achieved the required minimum threshold. This will be completely fair, because everyone has to do it in the same way. The proposed provider of the BCAT, Pearson Vue, has systems in place to cater for students requiring adjustments. Throughout the pilots, detailed studies and comparisons were made according to gender, primary language, age, ethnic origin and disability. The test is not aimed (as other such tests, eg LNAT) at selecting the best candidates from a large pool of applicants, but as a means of identifying those unsuitable to do the course, at a threshold level, and whose presence on the course adversely impacts on the learning experience of other students. **Conclusions** This policy is designed to eliminate the possibility of discrimination, to promote equality and the fostering of good relations between diverse groups. There has been a conscious and deliberate approach to considering the implications for the equality objectives before the decision was made. **NB** *For full details, please see the Aptitude Test*

	consultation paper (which includes this Impact Assessment as an Appendix).

Do you consider the policy to have an adverse impact on equality?

Gender equality Yes ☒ **No** ☑

Race equality Yes ☑ **No** ☒

Disability equality Yes ☒ **No** ☑

If yes, is there any evidence to support this?	**Race** There are some links between race and language, although this is not always the case. It might therefore be argued that imposing an aptitude test could disadvantage those candidates with poor skills, particularly those in the English language. Such skills are however vital to do the course, and train for the Bar. In addition, the requirement could be argued as potentially discriminatory against BME or non UK students because some skills that will be incorporated in the test, specifically English language skills are sometimes related to such backgrounds. However, this matter was thoroughly investigated by the BVC Review Working Group 2007-08. It seems evident that applying the same test for all applicants is the fairest method. The thoroughness of the two major pilots (over two academic years) has also rigorously examined such concepts. In fact, evidence shows that Asian students tend to actually perform less well on the course than their Test scores would predict. This suggests that if there is bias in the test it is against the higher scoring White group rather than the lower scoring Asian group.
	Gender There are no gender implications. Although in wider studies of Aptitude testing (for example the Dewberry Report) some small differences have been found for gender. However, the pilots carried out by the BSB show that no significant differences for gender were found for the BCAT.
	Disability Applicants with physical or other disabilities (eg dyslexia) are catered for by Pearson Vue in their test centres.
If no, are your reasons for this?	The change will not have any impact on race or gender equality, as the rule applies to all applicants in exactly

	the same way. Reasonable adjustments will be made for those with disabilities. There is no identifiable impact in relation to the age of candidates.

What are the (potential) barriers to equality arising from this policy? What evidence supports the existence of such barriers?	**Race** The change will not have any impact on race equality, as the rule applies to all applicants in exactly the same way. The final report indicated that no substantial adverse impact was found using a range of low to moderate cut scores for gender, primary, language, age, disability or whether the first degree was from a Russell Group institution. The largest impact was for ethnic groups with students from minority groups being somewhat less successful than the white group overall. However, as mentioned above, the results show that some BME categories actually perform better on the Test than on the course, so thus would not be unfairly excluded. The Test would actually operate in favour of this group for example. A low cut score will additionally minimize the impact on different ethnic groups.
	Gender The change will not have any impact on gender equality, as the rule applies to all applicants in exactly the same way. There may be a potential barrier for people who are not familiar with taking on-line tests but no relation to gender has been identified. This may impact more significantly on older people. Practice tests will be available to minimise any possible disadvantages.
	Disability Approximately 6% of students annually disclose that they have a disability although the true figure may be higher. The largest disability group is students with dyslexia. The introduction of the Aptitude Test will therefore need to include the provision of reasonable adjustments to ensure that students who are disabled in anyway are not disadvantaged. Pearson Vue, the appointed BCAT Provider has a clear policy on providing reasonable adjustments to candidates with disabilities.

	Candidates with a disability or condition which might require special arrangements are asked to discuss the matter with their test centre as soon as possible. Cases are considered individually and students will need to provide medical certification. Test centres may need a period of time to put arrangements into place. Measures would include such adjustments as enhancement in font size for those with visual impairments, extra time for those with difficulties such as dyslexia, wheel chair access and so on.

Stage 4 – Action Planning

Recommendations and actions required to reduce/remove barrier	**Person Responsible**	**Deadline**
The BSB will carry out a series of pilots of the BCAT in order to review its efficacy as a predictor of suitability for the Course. An independent consultant is, and will continue to be, appointed for this purpose.	BSB Head of Education Standards and the Independent Consultant	November 2011
Ensure an appropriate cut-score that will minimise the impact on different ethnic groups	BSB Head of Education Standards and Education team	From April 2012
Ensure that appropriate reasonable adjustments are made for candidates taking the test, where needed	BSB Head of Education Standards and Education team in conjunction with the Test Providers	From April 2012
Publicise the change to the entry requirement (once approved) to prospective students and the reasons for its introduction.	BSB Head of Education Standards and Education team in conjunction with Inns and Providers	From April 2012
Publicise the process for requesting reasonable adjustments for prospective students and emphasise timescales.	BSB Head of Education Standards and Education team in conjunction with Inns and Providers	From April 2012