The UKCAT-12 Study

Technical Report

Educational attainment, aptitude test performance, demographic and socio-economic contextual factors as predictors of first year outcome in twelve UK medical schools

IC McManus, C Dewberry, S Nicholson, J Dowell

Background

The following Technical Report was previously submitted to the UKCAT Consortium in March 2012 as a Confidential Document to inform the Consortium of progress in analyzing the UKCAT-12 datasets. The full report is now being released, unchanged, to coincide with and to supplement the publication in *BMC Medicine* of a group of three papers, two of which analyse the UKCAT-12 data in more detail.

The Report provides more details on some of the analyses and may be of use to those interested in UKCAT. It should be emphasized that the Report was a working document, that none of the conclusions were final, and indeed some of the conclusions and the methods on which they were based have been updated, in some cases as a result of minor errors having been found. The reader should therefore interpret the findings in that context and in conjunction with the three *BMC Medicine* papers.

The following page provides a brief summary of some of the findings of the Technical Report and the UKCAT-12 study, similar to that provided as a briefing to the UKCAT Consortium.

Confidentiality. The original document had ‘Confidential’ as a watermark across each page, and on the front page the statement, “This is a draft report prepared for discussion by UKCAT, which should not be published or discussed outside of UKCAT without permission of the UKCAT chair”. Since the document is now in the public domain, those confidentiality statements have been removed.

A brief summary of the UKCAT-12 paper and the Technical Report

Both the technical report and the UKCAT-12 paper represent findings from a prospective study of medical students in twelve UK medical schools, for whom UKCAT scores and other measures of educational attainment prior to entry to medical school were available, in relation to first year medical school examination results. Twelve UK medical schools, four of which were in Scotland, took part. The study population was almost 5000 students who had taken UKCAT in 2006-8, entered medical school in 2007-9, and took first year examinations in 2008-10. Collectively, this represents one of the largest studies of the relationship between entrance variables and medical school outcomes conducted to date.

The Predictor Variables were; Prior Educational Attainment (PEA: a composite measure, involving Higher and Advanced Higher results for students from Scotland and A-level, AS-level and GCSE results for students from England); '3 Best A-levels/5 Best Highers, as used by most UK Medical Schools; and total and subscale UKCAT scores.

The Outcome variables were; actual scores on exams in Skills, Knowledge and a Combined score, standardised to z-scores, within each of the 12 medical schools and cohort; and 'Progression' data (how many students left, repeated the year, passed the year after resits, or passed first time).

Background variables included (but were not confined to) sex, ethnicity, school type, and 'age' (in the sense of mature as compared to school leaver candidates).

Results

No clear picture on progression as an outcome variable emerged. The results indicate that UKCAT has a small predictive effect for first year exams, which is somewhat larger for graduates. It is of similar magnitude to that offered by '3 best A-Levels/ 5 best higher' which most UK medical schools weight heavily. However overall Prior Educational Attainment is an appreciably stronger predictor of first year performance.

In terms of the background variables, there was underperformance in outcomes for males, non-whites, and candidates from high performing schools (whether state or private).

No differential picture emerged from analysis of the Sub-scales or from splitting the 'Combined' scores into skills and knowledge. Verbal reasoning had the highest relationship with the 'Combined' and 'Knowledge' scores, but the lowest with 'Skills'.

A key statistic is the incremental validity of UKCAT over A-levels. This is 0.057 over PEA, and 0.101 over '3 best A levels/5 best Highers'.

Conclusions

UKCAT is a significant predictor of performance in first year exams, equivalent to '3 Best A-levels', but weaker than combined 'PEA'. The published paper further confirms the current validity of using all the existing measures of educational attainment at selection decision-making. However in the common absence of these UKCAT offers small but significant incremental validity which is operationally valuable.

The UKCAT-12 Study

Technical Report

Chris McManus

University College London

Chris Dewberry

Birkbeck, London

Sandra Nicholson

Queen Mary, London

Jon Dowell

University of Dundee

Executive Summary

1. The UKCAT-12 study was set up by the UKCAT Board to assess the predictive validity and incremental validity of UKCAT in a large group of entrants to UK medical schools.
2. This report provides a summary of other studies which are known to have evaluated the use of ability/aptitude tests and measures of educational attainment as predictors of university outcome. An overview is provided in the main text, and a detailed list of studies known to us is in the Appendix.
3. A general conclusion of the review is that prior educational attainment predicts outcome fairly well, both at medical school and at universities in general. Aptitude/general cognitive ability tests perform much less well, correlations with outcome depending on the extent to which the measures directly assess scientific knowledge and attainment. UKCAT is a relatively 'pure' test of general cognitive ability which explicitly does not attempt to assess scientific knowledge.
4. The 4,811 students participating in UKCAT-12 entered the twelve collaborating medical schools in the years 2007 to 2009, having taken UKCAT in 2006 to 2008, and examination results on a continuous scale were available at the end of the first year at medical school (exams taken in the summer of 2008 to 2010).
5. Overall, 2.0% of students failed or left the medical school for other reasons; an additional 2.0% repeated their first year; 11.7% passed their first year after resitting exams; and the remaining 84.3% of students passed without resits. There were only minimal differences between those leaving the course for academic reasons (n=55) and those leaving for non-academic reasons (n=50).
6. Information was not available on assessments in later years at medical school. It may be that the predictive value of the various measures is different for the clinical rather than basic medical science years, or indeed for postgraduate training, although it is too soon to test such hypotheses.
7. A wide range of background measures was available, particularly including detailed measures of Educational Attainment at A-levels and at SQA Highers. In addition information was available on demography and social background, and contextual measures were available for the school attended (in England), as well as measures of socio-cultural environment (the English Indices of Deprivation).
8. Educational attainment and UKCAT scores differed between the twelve medical schools, as did many of the background measures. For most analyses, outcome measures, as well as measures of educational attainment and UKCAT, were standardised within medical schools and cohorts, meaning that in most cases results only apply relative to other students within a medical school year.
9. Conventional multivariate statistics, as well as multi-level modelling, allowed independent effects of the many correlated measures to be assessed, and in particular differences in predictive value (slope) across medical schools could also be evaluated.
10. Although multi-level modelling does allow for the possibility that measures behave differently in different medical schools, in practice we found no evidence for such variance. This means that the impact of differences in student characteristics was independent of the medical school they attended.

11. A key question in interpreting UKCAT is how it predicts in relation to the more conventional measure of prior educational attainment. Before therefore considering the predictive value of UKCAT a series of fairly extensive preliminary analyses of the various measures of educational attainment was required, enabling an overall score of educational attainment to be derived, which was comparable across A-levels and Highers.

12. Our composite measure of educational attainment was a major predictor of medical school outcome (r=.391). Conventional measures of best three A-levels or best five Highers showed much weaker prediction of outcome, because most students were at ceiling. However derived measures using information from all A-levels, AS-levels and GCSEs (and detailed Highers marks, as well as Advanced Highers in Scotland) showed there was much unused predictive information in educational attainment measures. Of particular interest is that Scottish Highers had more predictive value (r=.464) than A-level derived measures (r=.331).

13. A range of background measures related to medical school outcome *after* educational attainment and other background factors were taken into account. In particular: Female students performed better; White students performed better; and students from non-selective schools performed better. In addition mature students also performed better, although here the influence of prior educational attainment could not be taken into account as information was not provided.

14. UKCAT correlated .142 with medical school outcome, which is not particularly high, but is typical of other aptitude/ability tests which are not assessing specific scientific knowledge. The predictive validity was substantially higher in mature entrants to medical school (r=.252) than in non-mature entrants (r=.137).

15. The incremental validity of UKCAT, after taking educational achievement into account, was low (r=.048). Once again that value is fairly typical of other studies of aptitude/general cognitive ability tests which do not assess specific scientific knowledge. Because three best-A-levels and five best Highers had lower correlations with outcome, mainly because of ceiling effects, the incremental validities of UKCAT were somewhat higher, with a value of .102 against three best A-levels and .073 against five best Highers.

16. Of the four sub-scores of UKCAT, the only one accounting for unique variance in first year medical school performance was Verbal Reasoning. In particular, high Verbal Reasoning scores were associated with *better* performance on 'Theory' exams, but *worse* performance on 'Skills' exams. Since a detailed description of the content of Theory and Skills exams was not provided it is not possible to provide any clear explanation for this difference.

17. Construct-level validity was calculated both for A-levels and total UKCAT score, correcting correlations for measurement error (reliability), and restriction of range in the selected population of entrants. The construct-level validity of three best A-levels was **0.854** whereas that for UKCAT total score was **0.239**. For technical reasons it was difficult to calculate the incremental validity of UKCAT but it was small, and possibly even negative.

18. Taken overall, UKCAT has a relatively low predictive validity for first-year medical school examination in non-mature students, but it might provide useful additional information in mature students, particularly in the absence of adequate information on prior educational attainment.

19. The UKCAT-12 study, despite the relatively negative findings concerning UKCAT itself, shows the strength of large, collaborative studies across a large number of medical schools for assessing important theoretical and practical issues concerning the nature of medical student selection in the UK, and the process of medical training.

Introduction

The United Kingdom Clinical Aptitude Test, UKCAT, which was developed by a consortium of UK medical and dental schools[1], was first administered in the summer and autumn of 2006 to applicants for admission in October 2007[a]. Twelve UK medical schools have made available information on the medical school performance of students who had previously taken UKCAT between 2006 and 2008, and entered those medical schools from 2007 to 2009, finishing the first year of their studies from 2008 to 2010. This report, on *The UKCAT-12 study*, examines the predictive validity of UKCAT and previous educational attainment (A-levels and Scottish Highers), in relation to first year medical school performance.

Aims and objectives of UKCAT. The 2006 UKCAT Annual Report described how the objectives of the new test were to improve the fairness and objectivity of selection for clinical subjects, and particularly to help differentiate between candidates at the upper end of the scale of academic ability, especially since many candidates achieved a maximum of three As at A-level, which might depend "more on the calibre and resources of candidates' schools than on the candidates' own abilities" [p.10][1]. There was also a broader concern that the abilities being tested by exams "might not have been entirely appropriate as a way to select students for the clinical professions" (ibid, p.10), and this issue is being addressed with the development of non-cognitive tests which are not part of the present analysis.

A major concern for UKCAT was to provide a test which,

> "allowed all candidates to compete on equal terms, irrespective of their educational background. This meant that we should concentrate on measuring aptitude, rather than knowledge: schools that require a measure of educational attainment should use another test (such as A-levels) alongside the UKCAT. [As a result…] the UKCAT tests a wide range of mental processes, and we believe this approach allows us to be fair to candidates from all backgrounds…" [p.12][1].

In consequence, UKCAT has four separate scores of different cognitive abilities, entitled *Abstract Reasoning*, *Quantitative Reasoning*, *Verbal reasoning* and *Decision Analysis*. The latter test was specially prepared for UKCAT, whereas the other three were based on existing test items developed by Pearson-Vue. One of the aims of UKCAT was to assess whether, "one section is a particularly good predictor of progress in medical school" (p.16), while taking into account the possibility that, "some schools may wish to pay particular attention to the specific aptitudes being tested" (p.16) (and by implication, some subtests may predict better in some medical schools than others).

Cognitive Ability, Aptitude, and Attainment UKCAT was specifically designed as a test of cognitive aptitude. Tests of cognitive aptitude assume that performance in a given role or job (or a set of roles and jobs) is associated with a specific and identifiable combination of cognitive and intellectual abilities. Typically these abilities include verbal, numerical and/or spatial ability.
It should be noted that in the last 20 years a large body of data from aptitude tests has been analysed[2-5], those analyses producing three findings of critical importance in relation to the assumptions underlying aptitude testing. First, factor analyses of large numbers of aptitude tests have revealed that one overarching cognitive ability factor, usually referred to as *g*, accounts for most of the correlation between different tests [6-8]. Second, this *g* factor emerges strongly across different test batteries, different racial, cultural, ethnic and nationality groups, and irrespective of the statistical technique (i.e. type of factor extraction) used [9]. Third, when information about specific combinations of intellectual aptitudes are added to information about general cognitive ability, there is little or no increase in the accuracy with which the job performance or training

[a] Note that numeric references refer to formal citations at the end of the report, whereas literal references refer to footnotes at the bottom of the page.

performance of people can be predicted [4,7,8,10-12]. One implication of these findings is that all sub-scales of tests of intellectual aptitude primarily measure general cognitive ability (g), and therefore that responses on these scales are likely to be strongly correlated.

It should also be noted that there is an important difference between **tests of cognitive ability**, such as UKCAT which are designed to examine a person's ability to solve novel problems, and **tests of attainment**[13]. Whereas tests of general cognitive ability are concerned with examining the degree of intellectual ability which someone can bring to solve novel problems, tests of attainment are designed to measure how well someone performs at a task requiring areas of knowledge or skill accumulated over a programme of education or training, or over life more generally. Performance at attainment tests is likely to be associated not only with general cognitive ability, but also with a variety of other factors including the extent to which someone has access to relevant learning material, how well they have been taught or trained, the degree to which they are interested in the material, and their level of intrinsic and extrinsic motivation to learn about the material.

Tests of attainment are sometimes included in aptitude tests. An example is BMAT, which incorporates a test of attainment measuring scientific knowledge. Scientific knowledge is not specifically tested in UKCAT on the basis that, "scientific knowledge is already tested by school-leaving exams, which have been shown to have a degree of correlation with performance on medical courses" [p.11][1] (and indeed one of us showed that A-levels, but not tests of general cognitive ability, predicted career progression both at medical school, and up to 17 years later[13].

To summarise, aptitude tests usually contain measures of specific forms of cognitive ability and sometimes also tests of attainment.

Outcome measure. In order to be fully evaluated, any selection test has to have a specific outcome against which its predictive validity can be assessed. The aim of UKCAT is stated as,

> "to select students who will perform well in medical … school and who will eventually make good doctors … The definition of a good clinical practitioner may be (justifiably) nebulous, and in the first instance we aim to establish whether the UKCAT can identify candidates who will fail during their undergraduate studies. […The aim is therefore to identify] those who are more likely to qualify and those who are more likely to perform well once they start practising their profession" (p.13).

A key phrase in that aim, for present purposes, is the identification of those, "who *will fail during their undergraduate studies*". The present study only considers data on the first undergraduate year at medical school, but since there is a growing recognition that later university performance is generally predicted by poor earlier performance, performance in the first year of study, particularly weak or bad performance, is a good test of the success of a selection test. It is also the case that those who fail their first year at medical school, and leave the medical school, will certainly not become good doctors.

Previous studies of attainment and aptitude tests in medicine and higher education.

Before considering the empirical details of the UKCAT-12 study, it is worth briefly worth considering other studies which have evaluated aspects of the predictive validity of educational attainment and aptitude tests, both for medical education in particular, and for higher education more generally. The Appendix to this report provides brief descriptions of the studies of which we are aware, and where possible tries to report the findings as a conventional correlation. Without carrying out a formal meta-analysis, we can summarise those previous studies as weighted averages of the correlations.

Predictive value of educational attainment. In *the medical student studies* the correlations of outcome with educational attainment are (.30 [meta-analysis of 62 studies], .36, .35 and .37), with the only outlier being with dropout as the outcome variable (.12). A weighted combination of these correlations is **0.31**. For studies of *university students in general* the correlations are .32, .31, .39, .38 and .38, with the large HEFCE study being quite the largest; a weighted combination gives a value of **0.38**. That the correlation is lower in medical students may in part reflect a restriction of range since most medical students, particularly nowadays, have high grades near ceiling. The single estimate based on dropouts may not be entirely accurate, and anyway may well be lower than for results overall since dropout is often not for academic but for other reasons. Taken overall, there seems much justification for the comment made, by the authors of the HEFCE report, that,

> *"There is a long-standing belief that achievement at A-level bears little relationship to what happens once students start their degree courses. … [Some authors] create the false impression that 'A-levels do not matter', which is far from the case"[para 18, our emphasis][14].*

Predictive value of aptitude tests. Aptitude tests for selection to medical school, or for selection to university in general, are controversial. UKCAT in particular has been criticised, not least for an absence of evidence of predictive validity[15], and medical school aptitude tests in general have been so criticised[16]. There are however several studies of the use of aptitude tests for selection, both within medicine in particular and for university more generally, and a brief summary of the key studies and their findings is in the appendix. Putting together the various studies, and with the exception of MCAT (which is very much an attainment test), it becomes clear that aptitude tests generally predict outcome measures to some extent, a weighted estimate from the available studies being **.13** for medical students, and **.14** for university students in general. In studies excluding attainment tests, partial correlations are typically small and of the order of .05. The table below compares typical findings in the literature of the predictive validity of aptitude tests and attainment tests, and provides a benchmark against which the UKCAT-12 study can be compared:

Typical (weighted average) correlations in studies of predictive validity		
	Medical school	Higher education in general
Educational attainment tests	**.31**	**.38**
Aptitude tests	**.13**	**.14**

A note on statistical power. Some difficulties in assessing the predictive ability of aptitude tests, as well as some of the apparent differences between studies, may reflect a lack of statistical power. For a true correlation of 0.3, a sample size of 92 gives a 90% power of detecting a significant difference at the 5% level with a one-tailed test. In contrast for correlations of 0.20, 0.15, 0.10 and 0.05 one needs sample sizes of 211, 377, 853 and 3422. Many of the studies described in the appendix will therefore have been underpowered for finding statistically significant but small associations between aptitude tests and performance.

Aims of the analysis of the UKCAT-12 data.

The present analysis takes into account the aims which UKCAT set for itself, as well as various previous studies of aptitude tests (and the criticisms of those studies). It therefore looks at:

- The predictive validity of UKCAT for performance in the first year of medical school studies
- The incremental validity of UKCAT over and above existing measures of educational attainment, both GCSEs/AS-levels/A-levels and Scottish Highers/Advanced Highers.
- The specific predictive ability, with and without taking educational attainment into account, of the four subscales of UKCAT.
- An assessment of whether 'Theory 'and 'Skills' measures at medical school are predicted differently by attainment and aptitude measures.
- An assessment of whether the predictive validity of any of the measures is different in the twelve medical schools that have taken part in the study.
- An assessment of the true predictive validity of UKCAT and educational attainment after correcting the data for restriction of range and for attenuation due to measure' unreliability.

Important features of the present analysis are that the sample is large (nearly 5,000 students), it is diverse (the data being collected in twelve different medical schools) and it is extended over time (the data being collected across several years). This gives the current study high statistical power, and also makes it possible to compare medical schools, and to assess the degree to which the conclusions can be generalized across them. Although UKCAT is currently used in the selection of both medical and dental students, the current analysis is restricted to medical students[b].

The present UKCAT analysis is important and unusual in many ways, not least in that it is rare in UK medical education to have equivalent outcome measures based on nearly 5,000 students at a dozen medical schools. Such data provide an opportunity not only to assess the narrow aims of the project (in particular, how effective is UKCAT?), but also to assess a far broader range of issues concerning the extent to which of a wide range ofvariables, educational, demographic, and social, have an influence on medical school outcome, particularly since the impact of many of those variables has not previously been examined properly.

Strategy of analysis.

A major interest for the present report is in the predictive validity of UKCAT, and in particular its *incremental validity*. Assessing incremental validity requires an assessment of the extent to which UKCAT provides additional predictive value over and above the prediction already provided by existing measures (of which A-levels, Scottish Highers and similar measures are the obvious ones). UKCAT may also be predicting because, in effect, it is acting as a surrogate for other measurable variables with which it is confounded and which themselves are predictive of outcome. A range of background measures therefore also need to be defined and included within the analyses. The strategy of the analysis is therefore:

1. ***The sample***. A clear, defined sample has to be identified, based on the data with which we have been provided. The sample here consists of all students at 12 UK medical schools for whom first-year performance measures and UKCAT scores are also available.
2. ***Outcome variables***. A clearly defined set of outcome (performance) measures needs to be identified, based on the data with which we have been provided. As far as possible this should

[b] Although it is also important to know whether the test is predictive of performance at dental schools, the present authors were only provided with data on medical school outcome, and therefore the analysis is restricted to that.To our knowledge there has not yet been any follow-up of dental students in the UKCAT consortium. Dental education has been much less studied than medical education, and while it is hoped that conclusions drawn in the study of medical students may be generalisable to dental students, that is not necessarily the case, and care should be adopted.

be a continuous measure of performance, standardised relative to each medical school and cohort of entry. As well as an overall measure, an attempt will be made to separate out what medical schools have called 'Theory' and 'Skills' measures.

3. ***Educational attainment measures***. Most university selection is based upon A-levels (and AS-levels and GCSEs) for students from England and Wales, and on Scottish Highers (and Advanced Highers) for students from Scotland. Although apparently straightforward, there are actually many subtleties in deriving clear measures of educational attainment, in part because candidates (and their schools and parents) choose which subjects to take and how many subjects to take, and it is far from clear that all are equally predictive. Medical schools have traditionally asked entrants for three science A-levels, with Chemistry being seen as particularly important, although there are no large-scale analyses of the predictive validity of individual subjects. A particular issue concerns General Studies A-level which is sometimes, but not always, treated as a full A-level. In recent years there have also been arguments, as A-levels have tended to reach a 'ceiling' of 3 A grades, that AS-levels or even GCSEs may have useful predictive value. In Scotland, Highers are the main basis for selection, although there is an interesting question concerning scoring, UCAS, and apparently most universities, treating A1 and A2 as equivalent, B3 and B4 as equivalent, and so on, although it is not clear whether they have equal predictive power. Likewise, many Scottish universities apparently ignore Advanced Highers, although they are taken by many applicants. Given this wide range of issues, important preliminary analyses are needed to assess a wide range of educational variables, separately for candidates from Scotland and elsewhere, and to assess which are good predictors of outcome, in order that they can be compared with UKCAT.
4. ***Influences on UKCAT and educational attainment***. UKCAT scores do not exist in a vacuum, but are produced by medical school applicants who have social backgrounds, go to particular types of schools, etc.. Before assessing the extent to which UKCAT predicts medical school outcome, it is desirable to assess the extent to which background measures, both individual and also contextual, can themselves relate to UKCAT scores. A similar set of concerns also apply to the measures of educational attainment, particularly those which are predictive of medical school outcome.

General comments and caveats

The medical schools providing data for this analysis have done so on the basis of strict anonymity. We have therefore removed any information which might readily identify institutions unless it is clearly impossible to do otherwise. We have also had no access to raw, non-anonymised data, and have had to accept the data as provided as being correct and accurate. There is good reason to believe it is so, but we can only take that on trust.

Data sources.

Data were provided by the Health Informatics Centre (HIC) at the University of Dundee, which is curating the data collected by UKCAT. The data were provided as a series of anonymised, encrypted files, with a randomised identification code for each student, which allowed the various datasets to be merged together.

HIC provided data collated from several separate sources:

1. ***Medical school outcome data.*** Information provided by medical schools on the performance of entrants in their first academic year.
2. ***UCAS data***. Information provided by UCAS (Universities and Colleges Admission Service) concerning:

   a. **Academic Attainment.** This consisted of either A-level, AS-level and GCSE results for students in England and Wales, or Scottish Highers, Advanced Highers, and Intermediates.

   b. **Background.** Information on sex, nationality, ethnicity, school type, etc.

3. ***Secondary school contextual measures.*** Information collected by the Department for Education on secondary school performance in England[c].

4. ***Socio-economic contextual measures.*** Postcodes for place of residence were used to link into the various measures collected as part of The English Indices of Deprivation[17], which are linked into small area census statistics and other sources and collectively known as Indices of Multiple Deprivation (IMD).

5. ***UKCAT.*** Two types of scores:

   a. Scores on the UKCAT tests

   b. Other data collected by UKCAT, including nationality and socio-economic background, and a contextual measure of how many individuals at a student's school had taken UKCAT.

Note that we describe data as **contextual** where it is not specifically describing an individual student, but rather an aggregate measure of secondary school attainment, socio-economic or other measures of the area where they live, etc. Although such measures should necessarily be treated with care since they are not measures at the level of the individual, contextual measures will in the future be provided routinely by UCAS, and they have also been shown to be of predictive value in assessing achievement at the BMAT aptitude test [18].

A note on the description of cohorts. The description of medical school cohorts can be confusing, and we have therefore followed the UCAS convention which refers to the year in which students enter university, as shown below:

Cohort Name	*UKCAT taken*	*Entered university*	*First year ends*
2007	Summer/autumn 2006	October 2007	June 2008
2008	Summer/autumn 2007	October 2008	June 2009
2009	Summer/autumn 2008	October 2009	June 2010

Where there is potential ambiguity we spell out which years are being referred to.

The primary database

Information is available on various numbers of individuals, not all sources having the same information. The **Primary Dataset** consists of all students for whom there are outcome measures provided by the medical school and a full set of UKCAT scores. Analyses will be restricted to these 4,811 individuals, unless otherwise stated, and descriptive statistics etc. will be based on the Primary Dataset. Of course many other measures, for a host of reasons, are missing within the Primary Dataset, and how those missing data are handled will be considered later.

Statistical methods and approaches.

Conventional statistical analyses used *IBM SPSS* 19.

[c] http://www.education.gov.uk/performancetables/16to18_10/england.shtml. No equivalent information was available for other countries of the United Kingdom.

The data being analysed here raise a number of technical statistical issues. In particular, as with all complex multivariate data, there are large numbers of **missing values** which need to be coped with. The data are also **multilevel**, so that differences between medical schools need to be dealt with. And finally there are a number **psychometric issues** which need considering, particularly to do with the interpretation of **correlations**.

1. ***Missing values***. For several different reasons large multivariate datasets often have a high proportion of missing values.. Some data are missing simply because students have not provided them (e.g. self-described ethnicity). Other data are missing for structural reasons (e.g. one cannot have a grade at A-level Physics if one has not taken the exam), and some data are missing because external databases only provide information on some groups of individuals (e.g. the English Indices of Deprivation are only available for those with English postcodes). There are many approaches in the statistical literature to the handling of missing data, and none is optimal. Some are clearly undesirable, and the most undesirable is 'listwise deletion', the default method of SPSS for multivariate statistics, which only retains cases if scores are available for *all* analysed variables For almost all realistic datasets this approach to missing data results in a considerable reduction in sample size and there is every reason to believe that that subset is biased. A common solution to missing data is to use 'mean substitution', missing values being replaced (or 'imputed') by the mean of the data for those cases which do have information. This has several problems, a) the means might be different for different subgroups, and b) those with missing data may be a biased subgroup (and for instance those taking A-level physics tend to have higher A-level results overall than those not taking it, making it inappropriate to substitute a population mean). Modern work in the handling of missing values[d] has developed down two routes, firstly using the EM algorithm to impute estimates for the missing values based on the full information available from other variables, particularly in the covariance matrix) and secondly, using the method known as 'multiple imputation', where the process of imputation is repeated a number of times with somewhat different results each time, so that a sensitivity analysis can be carried out. Although SPSS can currently carry out multiple imputation, for the present analysis we decided to use the EM approach, which in SPSS is carried out by the MVA function. In order to allow a comparison of the MVA approach and the more conventional mean substitution approach, for the analysis of A-levels and Scottish Highers, we provide correlation matrices and factor estimates using both methods (see tables 2 and 4),

2. ***Multilevel modelling***. Conventional multiple regression assumes that a dependent (Y) variable is a linear function of one or more independent (X) variables. That works well if all of the sample points (the individuals in the study) are independent of one another (as for instance, if one were looking at the relationship of weight (Y) to height (X) in a random sample of people. 'Independent' here can be construed as any one individual, on average, being no closer to any one of the other individuals in the sample.

   Often however, and particularly in education, individuals are clustered. Multilevel Modelling (MLM) was developed with educational contexts in mind, and for instance one may have a group of school-children who take a series of tests (tests are nested within individuals), the children are organised in classes (children are nested within classes), the classes are nested within schools, the schools are nested within educational authorities, the educational authorities are nested within countries (say, England, Wales, Scotland and Northern Ireland), and the countries are nested within the United Kingdom. The assumption of independence here breaks down. The children in a school are closer in some

[d] See www.missingdata.org.uk for information on some of the various methods, etc..

sense to the other children in their school than they are to children in another school. The analysis of social processes means that one often wishes to be able to decompose data in order to find the level at which processes work – do they depend on which class a child is in, which school they are in, and so on.

Multilevel models also take random effects into account. Most statistical models consider fixed effects. One might, for instance, wish to know if the relationship of weight to height is the same in men and women, i.e. whether there are sex differences. For practical purposes there are only two sexes, and hence as a factor sex is called 'fixed', and any study would sample both sexes. In contrast, there are very many school classes, and any study of schooling could, for instance, only look at a small, random sample of the total universe of school classes. School class is therefore a random effect rather than a fixed effect.

Multilevel models allow one to ask natural questions about effects which are random. In the present case, the 12 medical schools included in this study can be regarded as a sample of UK medical schools (and they are a small sample of all medical schools). A key practical questions in education is not only whether one variable predicts another (e.g. does UKCAT predict medical school outcome), but also whether the predictive value of UKCAT differs in some medical schools compared with others (because, for example, medical schools teach differently. MLM allows the fitting of multiple regression lines, so that one can ask, for instance, is there a difference between medical schools in the slopes or the intercepts of the regression lines, and is there a correlation between, say, those slopes and intercepts. MLM is the appropriate form of analysis for the UKCAT data as those are the key questions which need asking.

Multilevel models cannot be fitted using conventional statistical programs such as SPSS, and instead special-purpose software is required. Here we will use the program *MLwiN v2.24*, which is distributed by the University of Bristol as a part of an ESRC initiative.

3. ***Psychometrics and the interpretation of correlations.*** Mostly in this report we will describe correlations between variables, and in particular we will describe correlations between measures such as UKCAT scores, measures of Educational Attainment, and measures of achievement in the first year at medical school. A few comments on correlations and their interpretation are worth making:

   a. **Correlations.** Correlations take values in the range of -1 to +1, and in general the larger the absolute value of a correlation then the better something is predicted, with a value of zero meaning no relationship. The upper limit of 1 is both seductive and misleading, as any value less than one looks somehow far worse (it always begins with "0.") and there is a temptation to assume that a correlation of, say,.5, is only halfway there. That temptation is worse still if one wishes to ask how much variance is accounted for, by squaring correlation coefficients, as small numbers less than 1 become even smaller numbers when squared. Care must be taken then in considering the percentage of variance accounted for.

   b. **Reliability.** A much greater problem is that of reliability. No measures in social or biological sciences are perfectly reliable. Reliability is expressed as a correlation coefficient. For example, test-retest reliability is concerned with how well peoples' scores on a test at Time B can be predicted from their scores at Time A (and vice versa). Internal reliability (usually referred to as internal consistency) is a measure of the extent to which peoples' scores inter-correlate on all of the specific items measuring the construct of interest in a test. It is usually measured with Cronbach's alpha coefficient, and alpha is sensitive to both the magnitude of the inter-item correlations and the number of items. The greater these correlations, and the greater the number of items, the higher alpha will be. Examinations are far from perfectly reliable. Fortunately medical students take many exams, even in one year, and the final mark derived from them is probably fairly reliable (and a typical figure for grade point averages from a meta-analysis[19] is of the order of .84). If medical school exams were typically no more reliable than that, then no predictor could have a correlation of more than .84 with an outcome (performance) variable. Likewise, no aptitude test is perfectly reliable (and indeed some are very unreliable). As explained above internal reliability of tests increases as a function of the number of items they contain, , and so while the UKCAT total score has a relatively good internal reliability (about .87, as shown earlier), the subtests have much lower internal reliabilities, mostly from about .55 to about .80, with a median of about .68. The reliability of predictors also sets limits on how high correlation might be.

c. **Range restriction**. Reliability coefficients, as well as correlations between tests and outcomes, are heavily dependent upon the range of scores present in a population. The narrower the range of scores (usually expressed as the standard deviation), the lower will be the reliability and the lower the correlation with an outcome measure (and that is a particular problem with, for example, postgraduate examinations[20]). A thought experiment shows why this is so; if an analysis were restricted only to individuals scoring exactly the same on a test then there would be no variance, and the correlation, in effect would be zero. Populations which have been selected inevitably have a lower range of scores than those who have not been, such as applicants in general, since selection inevitably picks out those with the highest scores on various attributes. Such factors tend to attenuate the 'true' correlations between selection methods and performance

d. **Correction for range restriction and reliability**. For the majority of the report we will be concerned with raw correlations, and the provisos about reliability and range restriction must be taken into account in interpreting them, particularly when the correlations are small. Sometimes those can look very small, too small to be of any practical significance[e]. Later in the report we will take reliability and restriction of range into account, in order to interpret the impact of the correlations in the broader population of medical school applicants, as opposed to the much narrower, more restricted population of medical school entrants.

4. ***Medical school outcome measures***. The number of medical schools using UKCAT in 2007, 2008 and 2009 was 23, 25 and 26. Data were available for students at a total of twelve UK medical schools, although data was not available for all schools in all cohorts, there being data from 11, 11 and 9 schools for the 2007, 2008 and 2009 cohorts, with a total of 1661, 1710 and 1440 students in the three cohorts. The total number of students at each school in the study varied from 87 to 945 (median = 335, mean=401, SD=243). Four of the schools were from Scotland and eight from the rest of the UK, and it is necessary to distinguish Scottish from other schools (due to most of their applicants and entrants having Scottish Highers rather than A-levels as entry qualifications, and the different types of examination potentially behave differently from one another). Medical schools have been given random identification codes, with schools 1 to 4 being in Scotland, and 11 to 18 in the rest of the UK. Medical schools were asked to provide a range of information on each student (see below), but not all information was available for all

[e] . For those who worry that correlations of .1, .2, or even .3 look very small, they would do well to remember, as Rosenthal has pointed out[21], that a very large study of the benefits of low dose aspirin on health, with 22,071 participants, had an effect size of r=.034.

students. Information was collected from medical schools by the UKCAT Central Office, and provided to the research team as anonymised files. Information was provided on:

1. Date of starting the course;
2. Outcome of the first year (including having to resit exams, having to repeat the first year, or leaving the course).
3. Rank order of students within year.
4. Percentage marks in examinations. Marks were divided into 'Theory' and 'Skills' exams, although further information was not available on which examinations were included under these headings at each school, and not all schools provided information on the separate components. One school did not provide percentage marks, but only gave information on the overall outcome of the first year.

It should be noted that the data provided are not for all candidates in the year in the medical school, but for all candidates in the year for whom UKCAT results are available, which may not be the same. No information is available on those who are in each year but for whatever reason did not take UKCAT. Information on ranks is also difficult to interpret, partly because ranks are not equal interval, and partly because they cannot be rescaled easily to percentiles, since the total number in the year is not always known.

An ideal outcome measure is continuous, with intervals being equal. Where possible, therefore, percentage scores have been used in the first instance. The procedure for calculating the overall outcome measure was therefore:

1. For all schools it was possible to use information about failure, resits, etc., to create a four-point scale, with steps "Passed all first time" (4), "Passed after resit(s)" (3), "Repeat first year" (2), and "Left medical school" (1). This variable is referred to as *OutcomeFirstYear4pt*.
2. For reasons to become apparent below, *OutcomeFirstYear4pt* , within medical schools and cohorts, was converted to a normal score, *zOutcomeFirstYear4pt,* based on the mean mark of the cumulative mark under the normal distribution[f].
3. Some schools, but not all, had provided an overall percentage score, as well as percentage scores for theory and skills exams. Where possible an overall percentage score was chosen for a medical school which combined theory and skills marks. If only separate skills and theory marks were provided, the theory mark was used (since there was no indication of the relative weighting of the theory and skills marks). If there was no overall, theory or skills mark available (in a few individual cases at a number of schools, and overall in one particular medical school), *zOutcomeFirstYear4pt* was used as a substitute for the overall mark. This means that a z score, within medical schools and years, is available for all students at all schools, which is called *OverallMark*.
4. Where possible, marks were calculated specifically for "Theory exams" and "Skills exams", which are called *TheoryMark* and *SkillsMark*.

For the 4,811 students in the Primary Database, continuous outcome marks were available in 4510 cases, and *zOutcomeFirstYear4pt* was used as a substitute in the remaining 301 cases, these cases mostly being for the one medical school not providing continuous measures, and the small remainder consisting of sporadic cases at other medical schools. The table below shows *OutcomeFirstYear4pt* and the histogram in figure 1 shows the distributions of *OverallMark :*

[f] This was done using the Rank Cases command in SPSS, with Rank Types set as 'Normal Scores'.

OutcomeFirstYear4pt	N	%
Failed / left medical school	96	2.0%
Repeated first year	94	2.0%
Passed first year after resits	565	11.7%
Passed without resits	4056	84.3%
Total	4811	100%

Pure theory marks and skills marks were less easy to calculate, and were only available reliably in 2075 and 3184 cases respectively. Figure 1 shows histograms of the scores. In the 2073 students for whom both marks were available, the Pearson Correlation was 0.566, and a scattergram of the relationship is also shown in figure 1.

5. *UKCAT measures of ability*

The UKCAT assessment was first administered in 2006, and results described here are from the first three years of usage of the test (i.e. tests administered in 2006, 2007 and 2007 to applicants for admission in 2007, 2008 and 2009). The Annual Reports on the test provide much background information on the tests [1,22,23], and in addition the present report sometimes also draws on the unpublished technical reports on each administration [24-26].

The cognitive components of UKCAT had a standard format across the three years[g]. The table below shows the number of scored and pre-test items on each measure, along with the internal reliability for each year the test was administered. There appears to be no information on internal reliability provided for the total score for the 2006 (entry 2007) administration, but it can be assumed to be very similar to the value calculated for the subsequent two years, given the similar internal reliabilities for the sub-scores.

		Reliabilities		
Test	Items (Scored+Pretest)	2006 (entry 2007)	2007 (entry 2008)	2008 (entry 2009)
Verbal Reasoning (VR)	40+4	**.74**	**.66**	**.65**
Quantitative Reasoning (QR)	36+4	**.71**	**.76**	**.61**
Abstract Reasoning (AR)	60+5	**.86**	**.82**	**.79**
Decision Analysis (DA)	26+0	**.58**	**.56**	**.58**
Total Score	162+13	**-**	**.87**	**.86**

The main analyses primarily consider the UKCAT total score, but consideration is also given to the sub-scores.

In considering the UKCAT measures, it should be remembered that for the present analyses we only were provided with data on those applicants who entered medical school, and we have no information on the total group of applicants. The correction of correlations for restriction of range, etc., therefore has to take into account information from external sources, particularly the Annual Reports.

[g] Non-cognitive test were introduced from the 2007 round of testing (2008 entry), but data on those tests were not made available for the present analysis, which concentrated entirely on the cognitive tests.

6. ***UKCAT scores by cohort***. Mean UKCAT total scores rose across the cohorts (2488, 2540, 2582 for 2007-2009; p<.001). The mean scores of all applicants (from the Annual Reports) also rose across the same years (2407, 2430, 2457), albeit not at such a high rate. Remembering that statistical equating based on Item Response Theory is used to ensure that scaled scores are on an equivalent scale and hence are comparable across years, then it appears a) that somewhat better candidates took UKCAT in 2008 than 2006, and b) entrants to medical school had higher scores in 2008 than 2006, even after taking into account the increasing scores of applicants in general (and a possible explanation for that may be that medical schools used UKCAT for selection more strongly for 2009 entrants than for 2007 entrants). The UKCAT total scores also had a smaller standard deviation in the medical students (211, 195 and 201) compared with the applicants (259, 255, 268), such restriction of range being typical of tests which are either used themselves in selection or which covary with measures used in selection.

7. ***UKCAT scores by medical school***. Medical schools showed highly significant differences on the UKCAT total scores of candidates entering them (Oneway ANOVA, p<.001), as well as on the subscores. No further analysis will be presented here except to emphasise that since the main thrust of the analysis is on *how UKCAT predicts <u>within</u> medical schools*, most of the remaining analyses in the study will be restricted to UKCAT scores standardardised as z-scores within medical schools and cohorts. That also makes it easier to interpret results, particularly in the *MLwiN* analyses.

8. ***Other measures describing the taking of UKCAT***. As well as performance measures, a number of other measures concerning UKCAT were also collected:

    1. *Date of taking UKCAT**.*** Candidates choose when to take UKCAT. Those taking the exam earlier may be more conscientious, confident or forward-thinking, whereas those taking it late may have prepared for longer. Two variables were calculated: **UKCATdayOfTaking**, the number of days after the opening date on which a candidate took the test (lower numbers therefore mean the test was taken earlier), and **UKCATdayOfTakingPercentileRank**, which is the same measure as before but expressed as a percentile rank (to allow for differences across years), lower scores indicating that the test is taken earlier.

    2. *Numbers of items not answered.* Counts were provided of the numbers of items answered correctly, incorrectly and not answered. The former cannot be readily converted into scaled scores. However the latter, which we call **UKCATskipped** is a useful measure of whether candidates reached the end of the test or ran out of time. Overall the median number of items omitted by the entrants was 4 (mean = 7.7, SD, 9.8, range 0 – 111, quartiles 0 and 11), with 25.9% of candidates omitting no items. In general, as in any multiple-choice test without negative marking, candidates will always do better on average by answering all items, even if that involves guessing. Not surprisingly there is therefore a reasonably large negative correlation between number of items skipped and total score attained (r= -.360, p<.001).

    3. *UKCAT time extension.* Candidates with special needs can request a time extension when taking the test, indicated by a variable **UKCATexamSeriesCode**. Overall 1.93% of medical school entrants had a time extension. These candidates performed significantly *better* than candidates under standard conditions (mean z-score = .213, SD 1.04 compared with mean = -.004, SD = .996 for standard conditions, p-= .038). The group with extended time also had skipped fewer items (mean 7.00 vs 7.67, median 2 vs 4; 33.30% vs 25.8% omitting no items; Mann-Whitney U-test, p=.024). However regressing total score on **UKCATexamSeriesCode**, after taking **UKCATskipped** into account, did not reduce the significance of the effect of having a time extension. There were no other differences between the two groups on UKCAT background variables.

4. *UKCAT school experience*. In their careful study of factors influencing performance on BMAT, Emery et al [18] found that candidates taking the exam who had come from secondary schools with more experience of BMAT (number of candidates per year taking the test), performed rather better, after taking other factors into account. A similar measure was therefore calculated for UKCAT, **UKCATcandPerSchool**, which is *a contextual variable*, and consists of the number of students per school per year who have taken UKCAT since its inception. Considering 4022 individual candidates with information on schools[h], a mean of 16.3 candidates from their secondary school had taken UKCAT, with a range of 1 to 89, median = 11, quartiles of 5 and 22, and SD of 16.1.

Analysis of which background factors affect UKCAT scores overall will be deferred until the full range of measures has been described.

Educational qualifications (A-levels, Highers, etc).

Analysis of educational qualifications is always complicated, not least because different candidates take different examinations with different structures and choose to study different subjects. The current data are more complicated still since information is available on A-levels, AS-levels and GCSEs in England and Wales (E&W) , and on Scottish Highers and Advanced Highers in Scotland. Only a tiny handful of candidates took both E&W and Scottish qualifications, and they will not be considered further.

Educational qualifications were provided to HIC by UCAS, and there are differences between cohorts. UCAS reports a very wide range of qualifications (see the online description of the tariffs[i]), and for the 2007 and 2008 entrants allowed a maximum of 10 qualifications, with that number increasing to 20 for 2009. It is possible therefore that occasional qualifications held by the 2007 entrants have been omitted by UCAS[j].

Extracting summary statistics for the educational qualifications is far from straightforward, with many possible options. It should be said straightaway that we have specifically not used the UCAS tariff, which is not really suitable for high-achieving medical students, not least because it incorporates a range of qualifications which are not necessarily appropriate[k]. We have chosen therefore to use the following measures:

1. ***A-levels (Advanced levels).*** A-level grades are scored on the basis that A=10, B=8, C=6, D=4, E=2, Else=0. A* grades at A-level were not awarded during the period of the present study. Many candidates had no A-level grades, and a few candidates had only one or two, and both groups were omitted from the calculations. The fourteen specific measures calculated were:

    a. **Alevel_number_total.** 2764 candidates had at least three non-General Studies A-level grades, those with 3,4,5,6+ A-level grades consisting of 58.2%, 36.6%, 3.8%, and 1.4% of this group.

[h] Only actual schools were included in this calculation (i.e. with continuous education from age 11 to age 18), and further education institutions, etc, were excluded.

[i] http://www.ucas.com/he_staff/quals/ucas_tariff/tariff

[j] In an initial exploration we found that the total reported number of reported A-levels and AS-levels in the 2007 cohort seemed to artificially capped at 10, whereas there was no such limitation in the 2008 and 2009 cohorts, where the maximum numbers of reported qualifications were 14 and 15, well short of the possible number of 20 in the database. Examples of 15 qualifications were a candidate with 5 A-levels and 10 AS-levels, and another candidate with 8 Scottish Intermediates, 5 Highers and 2 Advanced Highers.

[k] Qualifications such as a Financial Services Diploma, Music Practicals, or the SQA Skills for Work.

b. **Alevel_Totalbest.** The summed grade for the three highest scoring A-levels. In 73.0% of cases this was the maximum score of 30 (i.e. AAA). However 21.3% of students had 28 points (AAB), 5.0% had 26 points (ABB), and 0.6% had 24 points, and four candidates had 20, 16, 16, and 10 points.

c. **Alevel_TotalPoints.** The total points achieved by a candidate for all of their A-levels. For those only taking 3 A-levels this was the same as the previous measure.

d. **Individual marks on Biology, Chemistry, Physics and Maths**. For each student, the grade was calculated for each individual subject[l] (**Alevels_highest_Biology**, **Alevels_highest_Chemistry, Alevels_highest_Physics, and Alevels_highest_Maths).** In addition four dummy variables were calculated (**A-levels_Taken_1_or_more_Biology**, **A-levels_Taken_1_or_more_Chemistry, A-levels_Taken_1_or_more_Physics, and A-levels_Taken_1_or_more_Maths),** scored as 1 if the subject had been taken and 0 if it had not. 95.7%, 99.1%, 24.8% and 63.3% of students had A-levels in Biology, Chemistry, Physics and Maths.

e. **Non-Science A-levels.** A variable **Alevels_Taken_1_or_more_NonScience** was calculated with a value of 1 if a student had one or more A-levels which were not the core sciences of Biology, Chemistry, Physics or Maths (or of course, General Studies). 49.9% of students had at least one such A-level. These A-levels were heterogeneous, including Arabic, English Literature, Economics, History, Psychology and Sociology, and the grade was not computed.

f. **General Studies**. Although not used for the total points measure, two variables were calculated for General Studies, **Alevels_highest_GeneralStudies** and **Alevels_Taken_1_or_more_GeneralStudies,** which are equivalent to those described above for core sciences. 26.0% of students had an A-level in General Studies, with 46.9% having an A grade.

2. ***AS-levels (Advanced Subsidiary levels).*** AS-level grades were scored on the same basis as A-levels, that A=10, B=8, C=6, D=4, E=2, Else=0. In particular:

   a. Fourteen measures were calculated in an almost exactly the same way as with A-levels, except that scores only counted if at least *four* AS-levels were taken, and the mean was reported for the best *four* AS-levels achieved. Variables are named in a similar way except that they begin **ASlevel…** rather than **Alevel…** .

   b. Rather fewer students had four or more AS-levels (n=1877) than had three or more A-levels (n= 2764). The reason for this is not clear, since AS-levels are, in some sense, a subsidiary part of A-levels[m]. AS-level grades showed rather more variability than A-levels, with, for instance, only 56.3% of candidates getting the maximum 40 points from their four best AS grades, compared with 73.0% of students being at ceiling on A-levels.

3. ***GCSE (General Certificate of Education).*** GCSE results were only available for the 2009 entry cohort, not having been collected by UCAS prior to that date. Scoring was, as far as possible, similar to that for A- and AS- levels.

   a. Grades for ordinary (single) GCSEs were scored as A*=6, A=5, B=4, C=3, D=2, E=1, else =0. Double Science and other double GCSEs were scored as A*A*=12, A*A=11, etc, and counted as two GCSEs taken. Only a tiny proportion of students with GCSEs had eight or

[l] In the few cases, which was mostly for maths, where there were two or more A-levels in related subjects, the higher/highest was used.

[m] We presume that the problem is something to do with the communication of examination results from Boards to UCAS, but we have not explored it further.

fewer grades, and therefore the overall grade was calculated as the sum of the *nine* best grades (counting double science as two separate GCSEs, etc.). Overall GCSE scores were available for 930 students, and they showed greater variability than A-levels or AS-levels, only 16.6% of students having the maximum of 54 points (equivalent to 9 A* GCSEs), with a mean of 49.0 and an SD of 4.7 (median = 50; quartiles 47 and 53).

b. Scores were also calculated for the individual core sciences of Biology, Physics, Chemistry and Maths, and in addition a score was calculated for Combined Science, which is regarded as a double GCSE, and hence scored 12, 11, etc. Combined Science was taken by 32.8% of students. An additional variable was also calculated for the number of non-core science subjects (**GCSE_Number_NonScience_Exams**). Variables are named in a similar way to A-levels, except that they begin **GCSE…** .

4. ***Scottish Highers***. The scoring of Scottish Highers (and Advanced Highers) is not straightforward. Much of the background to the qualifications, along with the difficulties posed for UCAS tariff scoring, can be found in the 2008 expert report to UCAS [27]. For present purposes, direct, absolute, comparability between A-levels and Highers (and Advanced Highers) is not required, but rather the relative attainment of students taking the different qualifications is needed. For scoring, therefore:

   a. Scottish Highers were in the first place scored in a broadly similar way to A-levels, with A=10, B=8, C=6 and D=4 (note that there is no E). A typical number of Highers in medical students is *five*, and therefore results were only included if students had *five* or more grades at Highers, with the five highest being summed.

   b. Other differences from A-levels are that there is no General Studies component, and almost all students will take a non-science Higher as a standard part of the exam. As a result measures for these components were not calculated. Scores for Highers are otherwise similar to Alevels in their nomenclature except that they begin **SQAhigher…** . Results for Scottish Highers were available for 769 students, with 72.4% gaining a maximum score of 50 points (based on the best five grades).

5. ***"Scottish Highers Plus"***. Although Scottish Highers and Advanced Highers are scored by UCAS and by most universities as A, B, C and D, the grades entered into the UCAS database are actually A1, A2, B3, B4, C5, C6, and D7. These results, with two bands at each grade, are presumably intended to be ordinal, and indeed are used in that way by some English universities[n] (although not, it would seem, Scottish universities). Using that as a model, Highers themselves were rescored (in what here are called "Scottish Highers Plus") as A1=10, A2=9, B3=8, B4=7, C5=6, C6=5, and D7=4. Scoring was otherwise as for Scottish Highers (above), with similar measures except that the variable names begin **SQAhigherPlus…** . Not surprisingly there is a wider range of scores, with only 19.9% of students gaining the maximum 50 points.

6. ***Scottish Advanced Highers***. The status of Advanced Highers is somewhat unclear, many Scottish universities not apparently requiring them or not treating them very seriously. In part this is an argument based in concerns about widening access, there being a worry that it is only selective schools which have the resources or provide the possibility of studying subjects at Advanced Higher level (although in the group of entrants, amongst 478 students from the state

[n] The supplementary information to be completed by applicants to Cambridge specifically requires bands within each grade(www.cam.ac.uk/admissions/undergraduate/publications/docs/saq.pdf). That such bands are used is shown at King's College London, where A2 is required at all Highers (www.kcl.ac.uk/prospectus/undergraduate/medicine/entryrequirements) and by the announcement by Churchill College, Cambridge that, from September 2009, its typical offer for Scottish students would be "A1, A1, A2" for Advanced Highers (http://www.chu.cam.ac.uk/admissions/undergraduates/typical_offers.php/).

sector, 93.1% had at least one Advanced Higher, compared with 81.8% of 237 students from the non-state sector). Overall, 573 students in the present survey had at least two Advanced Highers (i.e. 74.5% of the 769 students with Highers), and a further 108 had one Advanced Higher, and it therefore seemed worth looking at the predictive value of those results. Scoring was as for "Scottish Highers Plus" (i.e. A1=10, A2=9, B3=8, B4=7, C5=6, C6=5, and D7=4). Scores were calculated for individual core science subjects, and in addition the highest overall score was calculated. Only 22% of the 694 students with at least one Advanced Higher had the maximum of 10 points on their best Advanced Higher, the median grade being 9 (mean = 8.4, ASD=1.5, quartiles 8 to 9, range 4 to 10), and 22.6% having 7 or fewer points. As will be seen later, we also compared the proportions of students with Advanced Highers in those from selective and non-selective schools.

The confusion around Highers is shown in the summary table in the eighteenth edition of Richards *et al*'s *Learning Medicine*[28] describing the entry requirements for 31 UK universities for candidates with Scottish qualifications. Of the five Scottish medical courses, all specify either AAABB or AAAAB at Highers as a minimal requirement, with only one specifying Advanced Highers (ABB). In contrast, and despite relatively small numbers of entrants, on the other 26 courses, outside of Scotland, 14 specify particular requirements in terms of Highers (the mode being AAAAB), and 21 also specify requirements at Advanced Highers, 10 in terms of two Advanced Highers (BB, AB or AA), and 9 requiring three Advanced Highers (all requiring AAB at minimum, with one requiring AAA).

Background measures (Socio-economic status, etc).

A wide range of measures was available, both individual and contextual, to describe the social and educational background of students. These included:

1. ***Nationality, sex, age, and ethnicity.***
    a. *Nationality.* Nationality was based on the online information provided at the time students took the UKCAT test. Of the 4811students, 4598 (95.6%) were UK nationals, and of the remaining 213 (4.4%), 176 (3.7%) were EU/EEA nationals and 37 (0.8%) were from outside the EU/EEA.
    b. *Sex.* Sex was based on information provided by UCAS. Of the 4811 students, 2081 (43.3%) were male and 2730 (56.7%) were female.
    c. *Age.* Age was based on the student's stated age in years at the time of taking the UKCAT test[o]. Ages ranged from 17 to 45 (mode=18, mean = 19.55, SD=2.84, median = 19, quartiles = 18 and 20). Age was missing for 45 students. 28.9% of students were aged 21 or over, and 1.3% were aged 30 or over.
    d. *Ethnicity.* Ethnicity was based on the standard UCAS coding. Altogether students were in 23 different categories, including 69 for whom ethnicity was missing, 214 who were coded as Unknown, and 54 and 138 who gave 'Not given' in two different categories. On a simplified six category basis there were 3057 White, 577 Indian sub-continent, 223 Other Asian, 92 Black, 140 Mixed and 60 Other. Since ethnicity was not a primary interest of the analysis, the 4149 students with clear descriptions of ethnicity were divided, as in many other studies (see the review in Woolf et al[29]), into White (n=3,057, 73.7%) and Non-White (n=1,092, 26.3%), the variable being called **Ethnic2**.
2. ***Schooling***.

[o] It would have been better to have the standard UCAS age, which is based on the matriculation date of 1st October, but unfortunately it was not available in the dataset that was provided to us.

a. *School type* School type was based on information provided by UCAS. Of the 4811 students, 69 had missing information, 360 were in UCAS's 'Unknown' category, 219 were 'Apply Online UK', and 86 were 'Other'. Of 4077 students for whom information was available, 1941 (47.6%) were classified as coming from Selective Schools ('Grammar School' or 'Independent School'), and 2136 (52.4%) from non-Selective Schools ('Comprehensive School', 'Further/Higher education', 'Sixth Form Centre' and 'Sixth Form College'). There was also a separate measure of Selective Schooling, derived from the DFES contextual data for England (see next section). The overlap between the UCAS and DFES classifications was good, but not perfect. A definitive measure entitled **SelectiveSchool** was therefore created with a value of 1 if either UCAS or DFES data suggested a school was selective and otherwise a value of zero. Of the 4811 individuals in the Primary Database, 1986 (41.3%) had evidence of having attended a selective school.

b. *Contextual school measures* Contextual measures on schools were based on information collected by the Department for Education (DfE) at Key Stage 5 for the academic year 2010 (file created May 2011). Using school codes provided by UCAS, HIC merged the DfE data with UKCAT data. Overall 22 measures were available in the spreadsheets. Here we restrict ourselves to:

i) *DFESshrunkVA.* This is an average measure of value added between Key stages 4 and 5[p]. It is normed with a mean of 1000 across the country, but here the mean is 1005.2, the median is 1006, SD = 21.3, and range = 928 to 1119. The implication is that the schools from which medical students come are on average contributing more value at this stage. Information was only available for 2,561 students.

ii) *DFES.AvePointStudent.* This measure calculated the average points gained by each student across all of their examination entries. This contextual measure was available for 2,586 students, with a mean of 894, SD of 157, and range of 348 to 1366.

iii) *DFES.AvePointScore.* This measure is similar to the previous one except that the average is at the level of the examination entry, rather than being calculated at the level of the student. This contextual measure was available for 2582 students, with a mean of 228, SD=21.7, and range of 151 to 281. As a check on the validity of the contextual measures, it was shown that students who attended selective schools scored more highly on all three measures. **Warning**: It should be repeated, once again, that **care is needed in interpreting contextual measures**, since they are not descriptions of the performance of students, but descriptions of the performance of the schools in which they were educated.

3. *Socio-economic and social background*.

a. *Socio-economic classification.* Socio-economic classification (SEC) was based on the online information provided by students taking UKCAT, and used the abbreviated version of the self-coded questionnaire (NS-SEC) provided by *UK National Statistics*[q]. SEC was calculated separately for the two parents (if provided), and the higher SEC used for this analysis. Overall of 4,091 individuals with usable information, 3,740 (91.4%) were in SEC group 1, 105 (2.6%) in group 2, 146 (3.6%) in group 3, 38 (0.9%) in group 4, and 62 (1.5%) in group 5, where group 1 has the highest socio-economic status.

i) *Contextual measures of social background.* For students with postcodes in England, the set of contextual measures entitled *The English Indices of Deprivation* [17] were available. These are provided as an Access database, which HIC merged into the UKCAT datasets.

[p] "Shrunk" is a technical term used in multilevel modeling, which is used by the DFES in calculating these values.

[q] http://www.ons.gov.uk/ons/guide-method/classifications/archived-standard-classifications/ns-sec/self-coded-version-of-ns-sec/index.html

For convenience and to ensure anonymity, variables were recoded into deciles, with higher scores indicating less deprivation. As a check on the quality of these measures, they were correlated with SEC and Ethnic2, and all of the contextual measures showed lower scores (greater deprivation) in students from lower SEC groups (2 to 5), and in students who were non-White. The *English Indices of Deprivation* are complex, with multiple scales, although most of the indices correlate positively with one another. Clearly not all indicators are as meaningful as others in the context of the educational and social background of medical students.

ii) *IMDOverallQualityDecile* This is a weighted combination of seven indicators, shown as the next items. It is the single best indicator of deprivation.

iii) *IMD1IncomeDecile* This mainly consists of indicators of material deprivation. There are also two supplementary indices.

   (1) *IMDIDACIDecile.* Income Deprivation affecting Children.

   (2) *IMDIDAOPIDecile.* Income Deprivation affecting Older People.

iv) *IMD2EmploymentDecile.* Mainly concerned with measures concerning employment and unemployment.

v) *IMD3HealthDisabilitySkillsDecile.* Concerned with measures of physical and mental health.

vi) *IMD4EducationDecile.* Measures of education, skills and training. Divided into two sub-domains:

   (1) *IMDChildrenDecile.* Includes attainment at Key stages 2, 3 and 4, secondary school absence and post 16 education, as well as university entrance.

   (2) *IMDSkillsDecile.* Skills or their absence in adults.

vii) *IMD5HousingAndServicesDecile.* Physical and financial accessibility of housing and local services. Two sub-domains:

   (1) *IMDhousingGeographicalBarriersDecile.* Mainly in terms of road distances to key local services.

   (2) *IMDhousingWiderBarriersDecile.* Measures of lack of access in terms of affordability of houses, etc.

viii) *IMD6CrimeDecile* Based mainly on violence, burglary and theft.

ix) *IMD7LivingEnvironmentDecile.* Measures of the quality of the living environment, divided into two subscales:

   (1) *IMDLivingEnvironmentIndoorsDecile.* Poor quality of living accommodation, such as houses in poor condition and without central heating.

   (2) *IMDLivingEnvironmentOutdoorsDecile.* Measures of poor environment in terms of air quality, road traffic accidents, etc..

Preliminary analyses of Educational Attainment in students aged under 21

Before the main analyses are carried out, a series of preliminary analyses is required to determine which educational attainment measures to include in the main analyses. These analyses will be carried out separately for E&W and Scottish qualifications. As mentioned earlier, having appropriate educational measures is necessary in order to assess incremental validity of UKCAT.

1. ***A-levels, AS-levels and GCSEs***. A-levels are known to predict academic outcome in both medical students and students in general, and therefore it is to be expected that they will have predictive value here, although there is a potential problem due to restriction of range, a majority of students gaining AAA. The predictive value of AS-levels and GCSEs is not so clear, although it should be noted, as mentioned in the Appendix (below), that the University of Cambridge has argued for their use in selection, not least as they are available at the time when an applicant applies, rather than after selection has taken place [30]. The predictive value of other measures is not so clear, although earlier reports suggested that entrants not taking A-level Biology had problems in the early clinical years [31], and Chemistry is, of course, almost universally required in those applying for medical school. The impact of Biology and other science subjects therefore needs examining.

    a. **Analytic strategy**.

        i) ***Outcome/performance measure.*** For immediate purposes the analysis will be restricted to the overall medical school outcome/performance measure (**OverallMark**), which is a z-score within medical schools and cohorts

        ii) ***Age.*** As will be noted again below, almost all of the students with detailed A-level and other results are below the age of 21 (presumably reflecting the inability of UCAS to collate such data with applicant records[r]). These analyses are therefore *restricted to students with at least three A-levels and aged under 21 at the time of entry to medical school,* of whom there are 2,725.

        iii) ***Missing values.*** Many values are missing, partly because data were not provided (e.g. not all students had AS-level or GCSE results), and partly for structural reasons (e.g. if a candidate had not taken A-level Physics they could not have a grade for A-level Physics). For simple correlations, the number of valid cases, N, is indicated. For multivariate analyses, the EM algorithm in the MVA function in SPSS was used to impute values which were missing for structural or other reasons.

    b. **Simple correlates of OverallMark**. Table 1 shows correlations of the various measures at A-level, AS-level and GCSE with **OverallMark**. Broadly similar patterns are visible at all three levels. The numbers of exams taken are of little importance, although the scores on the N highest are strong predictors of **OverallMark** (and the N best scores is always a better predictor than a total points score). In general the choice of subjects taken has little predictive value, although there are statistically significant though very weak associations between OverallMark and taking GCSE Maths, taking A level Physics, and taking A and AS level General Studies. At GCSE there is no evidence that those taking Double Science do any better or worse than those not taking it (and doing single core sciences).

        With respect to grades obtained at GCSE, A- and AS-levels all of the core sciences predict outcome in all three of these school qualifications, as also does grade at General Studies. Discussion of this will be left until after the multivariate analyses.

    c. **Multivariate analyses**. Multiple regression was used to explore the significant correlations in Table 1, in a series of steps:

[r] Mr Tim Williams of UCAS (personal communication, 16th Jan 2012) has confirmed that UCAS has basic applicant data going back to 1996, basic tariff data going back to 2002, and A-level data going back to 2006. GCSE data have only been available since 2009, the most recent group being studied here. Information on A-levels of graduates would not therefore be available for most of the group of graduates being considered here. It should also be noted that UCAS says that it intended to keep all such data in perpetuity (despite some rumours to the contrary).

i) In Steps 1 to 3, the points from the N best grades for A-levels, AS-levels and GCSEs were entered in that order. All three were individually significant. On its own A-levels had a beta of .177, and hence also an R of .177. Adding AS-levels, increased R to .225, with betas of .100 and .158 for A- and AS-levels, and including GCSEs increased R to .262, with betas of .066, .084 and .166. Although AS-levels and GCSEs appear to be better predictors than A-levels, that in part is due to the narrow variance associated with A-levels. For A-levels on their own, b=.138 (i.e. an increase of .138 standard deviations on OverallMark for each point on the A-level score, which was linear over the range 30 to 24 points (i.e. AAA to BBB), so that the examination performance of first year medical school entrants with BBB was about 1.1 SDs below those with AAA, a robust effect were it to be extended down to CCC or even DDD.

ii) At step 4, the variables were entered for General Studies being taken and the grade at General Studies (and it should be remembered that General Studies was not included as one of the 'best' A-levels). There was a highly significant predictive effect of General Studies grade (beta=.358), but no effect of having taken General Studies (p=.731); the latter was therefore dropped from the model. R was now .373.

iii) At step 5, the variables for the grades at the Core Science subjects were entered. Somewhat surprisingly, all four were highly significant (Biology, beta=.074, p=.005; Chemistry, beta = .129, p<.001; Maths, beta=.108, p<.001; Physics, beta=.091, p<.001); it was also the case that best of N variables for A-levels, AS-levels and GCSEs all remained significant (p=.002, .049 and .004, as did grade at General Studies (p<.001). For this model, R=.387, accounting for 15.0% of variance in OverallScore.

iv) The findings of the multiple regression seem surprising since the natural assumption is that the best 3 A-levels, which don't anyway seem to have that much variance, would automatically include the variance from the four core sciences. A little further exploration was therefore carried out of the four core sciences:

   (1) Core sciences taken. Although a majority of students had taken three core sciences (56.5%), 13.5% had taken four core sciences, 29.5% only two core sciences, and 0.6% just one core science.

   (2) Grades at core sciences. Given that students had taken from one to four core sciences, four measures were calculated from them. These are shown below, along with their correlations with OverallMark:

      (a) Mean grade at core sciences (irrespective of the number); r=.199;

      (b) Total points at core sciences; r= .095;

      (c) Maximum grade at core sciences; r=.066;

      (d) Minimum grade at core sciences taken; r=.213.

v) Considering the results above, the most parsimonious picture is that the most important predictor of outcome is the *minimum core science* grade. Numbers of core sciences taken do not seem to matter in particular, but what does seem to matter, irrespective of grades in other non-science subjects, or indeed other core sciences, is that *a student has achieved a low mark on one science which they have chosen to study*. That low mark presumably shows a failure to grapple with that science, and since medicine is a melange of many different sciences, may also show a potential failure to grapple with medicine as a science.

d. **Strategy for using A-levels in the modelling of UKCAT prediction**. On the basis of the results described above, for students taking A-levels, outcome will be guided by the analyses above. For convenience it is best to have one overall index of educational achievement for comparison with UKCAT, and therefore a single measure was obtained using Principal Component Analysis (PCA) . The PCA was carried out using the *SPSS Factor* program, with only the first component being extracted. A score, **EducationalAttainment,** was calculated for each student, using the Regression Method in SPSS Factor. Eight variables were included in the analysis:

i) Alevel_TotalbestN ASlevel_TotalbestN

ii) GCSE_TotalbestN

iii) Alevel_highest_GeneralStudies

iv) Alevel_highest_Biology

v) Alevel_highest_Chemistry

vi) Alevel_highest_Maths

vii) Alevel_highest_Physics

viii) Factor analysis used a data set in which EM imputation was used to substitute for missing values. The eigenvalues were 3.61, 1.24, .91, .78, .63. .45..30 and .09, with the first being by far the largest, accounting for 45.1% of the variance, and suggesting a single common factor. Table 2 shows the loadings of the eight variables on the first principal component and all are large and positive. As well as missing values being imputed by the EM algorithm, they were also, for comparison, handled using Mean Substitution. Of the eigenvalues (2.35, 1.13, .98, .93, .88, .84, .73 and .18), the first was then rather lower, accounting for only 29.4% of the variance, and the loadings were far less even, measures which often had missing values having much lower loadings. The mean number of variables missing was 2.96 (SD=1.01, median=3, quartiles 2 and 4, range = 0 to 6). Mean Substitution is far less satisfactory, particularly when a high proportion of candidates is missing, as they are substituted at the mean, which reduces the overall variance. Also with mean substitution the mean itself not be a not good value to substitute if, say, those with missing values are a biased subset of the population, and that variable correlates with other variables. The EM algorithm takes cares of most such problems.

e. **Educational Attainment in relation to three best A-levels.** Although our composite measure of **Educational Attainment** has many statistical advantages, we are aware that it has the disadvantage that it is not easily computed by admissions tutors, candidates, etc. As well as using it we will also, therefore, report correlations with the three highest A-level grades attained (**Alevel_TotalBestN**), which correlates .690 (n=2725) with the EducationalAttainment measure. The correlation is not particularly high because 73.0% of students are at ceiling, with 30 points. In considering such "three best" (and "five best") measures it should be realised, particularly given the fact that both measures are mostly at ceiling, that it does not make sense to standardise the A-levels and Highers measures. In addition, there are highly significant differences in best A-level grades and best Highers grades between medical schools, which also makes standardisation within schools difficult. Given all these problems it is not surprising that correlations with best A-levels and Highers are consistently low (and often low), in comparison with the composite measure of Educational Attainment.

f. **Discussion of predictive value of A-levels, etc.** Although not the principle purpose of this study, the data on A-levels, AS-levels and GCSEs raise a number of important issues:.

i) Although A-level grades may appear to be at ceiling, there is still predictive variance present.

ii) That grade obtained in General Studies has additional predictive value is somewhat of a surprise, but therefore of particular interest.

iii) There is no one science that seems particularly to be predictive of outcome.

iv) The Core Sciences predict in a way that was not expected, the implication being that achievement at each contributes to medical schoolperformance.

v) A key predictor of (poor) examination performance in the first year of medical school is previous poor attainment in at least one core science.

2. ***Scottish Highers and Advanced Highers***. The preliminary analysis of Scottish Highers and Advanced Highers proceeds in a similar way to that of A-levels, etc. Far less has been written in the literature about the predictive value of Scottish examinations, and a key question is their similarity or otherwise to the E&W qualifications. A difference from the E&W qualifications is that whereas GCSEs, AS-levels and A-levels follow in a direct causal order, with the latter being the name qualification used for selection, for the Scottish qualifications it is Highers which are used for selection, 'HighersPlus' has been synthesised for present purposes, and Advanced Highers are taken by only a minority of Scottish candidates (overall), although they are actually taken by a majority of entrants to medical school, and even those admitted typically have only one or two Advanced. Nevertheless the analytic strategy can be broadly similar to that used for the E&W qualifications, with the Scottish exams providing an independent replication for the findings south of the border.

a. **Analytic strategy**.

i) **Outcome measure.** As with the E&W exams, the analysis is restricted to the overall outcome measure (**OverallMark**), which is a z-score within medical schools and cohorts.

ii) **Age.** As previously, almost all of the students with detailed Scottish results are below the age of 21 (presumably reflecting the situation at UCAS). Analyses are therefore *restricted to students with at least five Highers and aged under 21 at the time of entry to medical school,* of whom there are 716, rather fewer than for the A-level analyses.

iii) **Missing values**. A number of values are missing, partly because data were not provided (e.g. not all students had taken Advanced Highers), and partly for structural reasons (e.g. not all students had taken all subjects, and if they had not taken Highers Physics they could not have a grade for Highers Physics). For simple correlations the sample size is indicated. For multivariate analyses, the EM algorithm in the MVA function in SPSS was used to impute values which were missing for structural or other reasons.

iv) **Highers and Advanced Highers in selective and non-selective schools.** As mentioned previously, there is a concern in Scotland that Advanced Highers are only available for students from schools which are selective. Amongst the 715 students taking Highers, we therefore compared the 237 students from selective schools with the 478 from non-selective schools (based on UCAS's classification). Students from selective schools in fact had taken somewhat *fewer* Advanced Highers (Sel: mean=1.86, SD=1.11, nonSel: mean=2.08, SD.90, p=.005). Looking at students from selective (non-selective) schools, 18.1% (6.9%) had no Advanced Highers, 12.7%(14.2%) had one Advanced Higher, 36.7% (44.8%) had two, and 32.5% (34.1%) had three or more Advanced Highers. The proportions of selective (non-selective) school students with Advanced Highers in various subjects were: Biology 69.1% (61.8%, NS), Chemistry 85.6% (80.0%, NS), Maths

28.4% (36.6%, p=.042) and Physics 10.3% (21.6% p=.001). Amongst these entrants to medical school it seems therefore that Advanced Highers in general, and particularly in Maths and Physics, are *more* frequent amongst non-selective school students than those from selective schools.

b. **Simple correlates of OverallMark**. Table 3 shows correlations of the various measures at Highers, 'HighersPlus' and Advanced Highers with OverallMark. The picture is broadly similar for all three qualifications, and there is also much that is similar to Table 1 for A-levels, AS-levels and GCSEs. The number of exams taken is of little importance, except for Advanced Highers, which are taken by few, and those who tend to take more also do better at medical school. The scores on the N highest exams are strong predictors of OverallMark, with much higher correlations for 'HighersPlus' and particularly for Advanced Highers. In general the choice of subjects taken has little predictive value, with the interesting exception of Chemistry at Advanced Highers, those choosing to take it tending to do better at medical school. At all levels, grades in each of the core sciences predicts outcome, the sole exception being Chemistry at Highers, where most students have top marks. It is noteworthy that the correlation with OverallMark becomes progressively larger from Highers, to 'HighersPlus', to Advanced Highers. The final five rows show analyses of grades at the Core Sciences, as was developed during the discussion of the analysis of A-levels. The broad picture is very similar, the mean grade at the Core Sciences particularly predicting outcome, and as before the minimum grade achieved at the Core Sciences is a particular predictor of medical school outcome. Taken overall, this pattern of results is extremely similar to that reported earlier for A-levels, AS-levels and GCSEs, and gives confidence in both analyses. As was suggested before, although simple outcome measures such as total points at Highers are close to ceiling, there is an underlying richness in the measures which suggests that prediction of medical school outcome is still possible.

c. **Multivariate analyses**. As in the analysis of the E&W exams, multiple regression was used to explore the significant correlations in Table 3, in a series of steps. As before the analysis was on the EM imputed data.

i) In Steps 1 to 3, the points from the N best grades for Highers, 'HighersPlus' and Advanced Highers were entered in that order. Highers alone were barely significant, and stop being significant when the other two measures were entered. It has therefore been dropped from the model. 'HighersPlus' was significant when in the model on its own (beta = .113, p=.002), but stopped being significant when the best Advanced Highers was entered, and therefore best 'Highers Plus' was also dropped. The best Advanced Highers result was significant in its own right (beta = .339, p<.001), and continued to be until individual Advanced Highers marks were included in the model. It was therefore also dropped from the final model.

ii) At the next step, the individual Core Science marks at Highers were entered. Maths and Physics were not significant and were dropped. Biology was then significant but was no longer significant in later models, and therefore was also dropped. Finally, Chemistry was not significant, and was dropped. The Highers measures therefore had little predictive power, as might be expected from Table 3.

iii) The individual 'HighersPlus' subject measures were next included in the model. Chemistry and Maths were not significant and were dropped. Biology and Physics were significant individually (beta = .103, .100; p=.010 and .012). However neither was significant at the next step, when Advanced Highers were entered, and both were therefore dropped.

iv) In a similar way the individual Advanced Highers subject measures were entered. Maths and Physics were not significant, but Biology and Chemistry were each individually significant (beta = .135 and .338; p =.002 and p<.001). Together the two variables had a multiple R of .439, accounting for 19.3% of the variance.

v) As a check on the model, after Advanced Highers Biology and Chemistry were entered, all of the other Highers, 'HighersPlus' and Advanced Highers measures were allowed to enter with a stepwise forward entry. Only two reached significance, Chemistry grade at Highers, and having taken Maths at Highers. Although both were significant with p<.001, each had negative coefficients. Given that neither was significant in its own right, that they had negative coefficients, and that their inclusion *increased* the significance of the two Advanced Highers measures, it was concluded that the effects probably reflected a combination of multicollinearity and suppressor variables.

vi) In interpreting the regression model, with just the two measures predicting OverallMark, it should be remembered that many students had not taken Advanced Highers in Biology and/or Chemistry. However, missing values for these variables were imputed via the EM algorithm, which takes into account the relationships of these measures for those for whom data are present, to estimate what grades would have been attained were they to have been taken. In effect, therefore, all other variables in the model have, to some extent, been included in the choice of these two measures. In fact, of the 715 students, 333 (46.6%) had Advanced Highers in both Biology and Chemistry, 262 (36.6%) had Advanced Highers in one of them, and only 120 (16.8%) had Advanced Highers in neither of them (and 76 of these had no Advanced Highers, the remaining 44 having one (27), two (14) or three (3) Advanced Highers in other subjects.

vii) Overall, the analyses support the conclusions reached for A-levels, AS-levels and GCSEs. As before, and particularly for Highers and 'HighersPlus', a parsimonious picture is that an important predictor of outcome is the *minimum core science* grade. The picture in Scotland though is made more complex and more interesting by the presence of Advanced Highers, which are chosen to be taken by a smaller numbers of applicants. Advanced Highers are clearly much more difficult for medical students, and there is a wider range of marks, providing much more variance which is then correlating with medical school outcome. It is also of interest that two of the four core sciences, Biology and Chemistry, and particularly Chemistry, have the greatest predictor power. Chemistry has long been thought to be of particular importance in predicting performance at medical school, and these results seem to support the particular emphasis on this subject, as well as a requirement for Biology.

d. **Strategy for using Highers in the modelling of UKCAT prediction**. The analysis of Highers followed the same approach as that for A-levels. Using the results above, the summary variable was based on ten measures, described below. The Principal Component Analysis (PCA) was carried out using the SPSS Factor program, with only the first component being extracted. A score, **EducationalAttainment** was calculated for each student, using the Regression Method in SPSS Factor. The ten variables were:

i) SQAhigherPlus_TotalbestN   Total score for best SQAhigherPlus

ii) SQAhigherPlus_highest_Biology

iii) SQAhigherPlus_highest_Chemistry

iv) SQAhigherPlus_highest_Maths

v) SQAhigherPlus_highest_Physics

vi) SQAadvHigher_TotalbestN Score for best SQAadvHigher

vii) SQAadvHigher_highest_Biology

viii) SQAadvHigher_highest_Chemistry

ix) SQAadvHigher_highest_Maths

x) SQAadvHigher_highest_Physics

Using EM imputation for missing values, the factor analysis has eigenvalues of 5.20, 1.56,.84, .63, .51, .49, .28. .24, .16 and .08. The first is by far the largest, accounting for 52.0% of the variance, and suggesting a single common factor. Table 4 shows the loadings of the eight variables on the first principal component and all are large and positive. As well as missing values being imputed by the EM algorithm, they were also, for comparison, handled using Mean Substitution. Of the eigenvalues (3.63, 1.40, .98, .91, .85, .69, .56, .51, .27 and .21), the first was then rather lower, accounting for only 36.3% of the variance, and the loadings were far less even, measures which often had missing values having much lower loadings. The mean number of variables missing was 2.93 (SD=1.83, median=3, quartiles 2 and 3, range = 0 to 10). Mean Substitution is far less satisfactory, particularly when a high proportion of candidates is missing, as they are substituted at the mean, which reduces the overall variance. Also with mean substitution the mean itself is not a good value to substitute if, say, those with missing values are a biased subset of the population, and that variable correlates with other variables. The EM algorithm takes cares of such problems.

e. **EducationalAttainment in relation to five best Highers.** As with A-levels, although the composite measure of Educational Attainment has many statistical advantages, it is not straightforward to compute. We therefore also report correlations with the five highest Scottish Highers (on the conventional scale), (Highers_TotalBestN), which correlates .328 (n=715) with the EducationalAttainment measure. The correlation is relatively low in part because, as with A-levels, 72.4% of students are at ceiling with 50 points at Highers.

f. **Discussion of predictive value of Scottish Highers, etc.** As with A-levels, although the principle interest of this study is not in the use of Scottish Highers, etc., for predicting medical school outcomes, nevertheless the analyses just reported show clearly that Scottish Highers, and in particular scores based on the full information in them ('HighersPlus'), and even more with Advanced Highers, make good predictions of outcome at the first year of medical school. That is important practical information. Although not all universities, particularly in Scotland, seem to choose to use the additional information in Highers grades, and also in Advanced Highers, that is no reason to exclude themfor present purposes.

Predictive validity of A-levels and Highers in Medical schools in Scotland and E&W.

The scores for Educational Attainment based on A-levels and on Highers are on equivalent scales since, being factor scores, they are dimensionless, and hence each have a mean of zero and a standard deviation of 1. That means they can be combined as a single score, which is called **zEducationalAttainment** to indicate that it is a z-score. In addition the overall score for first year medical school performance has been rescaled so that it has a mean of zero and a standard deviation of one *within cohorts within medical schools* (since the interest throughout is on prediction within schools and cohorts, rather than on comparing schools and cohorts[s]). The

[s] It is worth saying that medical schools undoubtedly differ in the Educational Achievement scores of those entering them, using the measures based on both A-levels and Highers, but that is of no relevance here.

correlation of OverallMark is higher with the Highers-based score (r=.464, n=715) than with the A-levels-based score (r=.331, n=2717). The distribution of the Educational Attainment measures based on A-levels is shown in figure 2, along with a scattergram of the prediction of OverallMark. The red (straight) line shows a linear regression function, whereas the green (curved) line shows the lowess function. The difference between the two lines suggests that the predictive value of A-level based achievement may be greater (the line is steeper) at higher levels of achievement (as would be compatible with several other large-scale studies in higher education[32]). Figure 3 shows similar results for Scottish Highers, and the scales and size of the two scattergrams are identical, so that the difference in the slopes of the lines can more readily be compared. As with A-levels, it is also the case for Highers that the lowess curve suggests a greater predictive value at higher levels of achievement.

A noteworthy feature of both Figures 2 and 3 is that the loess curve is concave upwards (i.e. the predictive value of educational attainment is *higher* at higher levels of educational achievement, the slope being steeper). Curvilinearity was tested formally using a quadratic regression which was highly significant (p<.001) for both A-levels and Highers, the quadratic term increasing R from .331 to .343 for A-levels, and from .464 to .505 for Highers. Both effects are compatible with the 'more-is-better' rather than the 'good-enough' hypothesis[32].

The difference in predictive ability of A-levels and Highers can be explored further since, although almost all students taking Highers are at Scottish medical schools, the converse is not the case and there is a reasonable proportion of students with A-levels at Scottish medical schools. The table below shows the correlations of OverallScore with Educational Attainment (as well as equivalent calculations for best 3 A-levels/ 5 Highers).

Educational Attainment score	Medical school in England and Wales	Medical School in Scotland
A-levels etc taken	**r = .336**, n=2271, p<.001 (95% CI .299 to .360)	**r=.280**, n=446, p<.001 (95% CI=.192 to .357)
Highers etc taken	n/a (n=9)	**r=.471**, n=706, p<.001 (95% CI = .412 to .497)
3/5 Best A-levels/Highers	Medical school in England and Wales	Medical School in Scotland
3 best A-levels	**r = .187**, n=2271, p<.001 (95% CI = .147 - .227)	**r = .136**, n=446, p=.004 (95% CI = .045 to .227)
5 best Highers	n/a (n=9)	**r = .122**, n=706, p=.001 (95% CI = .049 to .483)

It seems clear from the confidence intervals that for the composite Educational Attainment measure, A-levels predict outcomes equally well in Scotland and E&W, implying that outcome measures in Scotland and in England and Wales are equivalent. The difference between Highers and A-levels seems therefore to be because Highers have a better predictive value, as can be seen in the Scottish medical schools (and that may reflect the fact, particularly with Advanced Highers, that many fewer students have achieved ceiling on the exams). The correlations with the best A-levels/Highers are far lower than for the composite measures, and show no real differences between the groups, but the measures are mostly at ceiling.

Relationship of social, demographic and other background measures to Educational Attainment, UKCAT performance, and Medical School outcome.

There is a wide range of background variables, and it is important, while assessing the primary concern of the predictive validity of UKCAT performance upon medical school outcome, that background variables, where significant, are known about and taken into account. In this section an overview will be taken of how the wide range of background measures relate to the three key measures.

1. ***Simple correlations.*** Table 5 shows simple Pearson correlations of each of the three key variables, as well as the 3 highest A-levels and 5 highest Highers, with 22 background variables, 10 of which are descriptions of individual students, and the remaining 12 are contextual measures and hence descriptions of the school or social environment of the student; once again, a warning must be given about the potential problems of interpreting such contextual variables. It should also be remembered that the N varies between the correlations, and many of the 22 variables correlate one with another, and no account of that is taken in Table 5. The three key variables on the columns are all aspects of ability and attainment, the first and third being measures of academic attainment at school and at medical school, and the middle one the UKCAT score, which is a measure of ability or aptitude. It will not be surprising if similar patterns of correlations occur for all of these three variables, although there may well also be important differences.

    a. ***Demographic measures.*** Discussion will mainly concern the correlations with the measure of Educational Attainment, UKCAT total, and OverallScore, and will mostly not consider best 3 A-levels and best 5 Highers.

        i) ***Ethnic origin.*** Being from an ethnic minority is strongly and consistently associated with poorer performance on all three key measures. The UKCAT is assessed routinely for DIF (Differential Item Functioning) in relation to ethnicity (and other factors such as sex and age [1,22,23]) and therefore the difference is probably real. The effect size for the medical school outcome measure is -0.33, which is compatible with the effect sizes found by Woolf et al [29] in their meta-analysis[t], the overall effect in undergraduates being -.42 (95% CI -.49 to -.35). The effect sizes for Educational Attainment and UKCAT score are -.21[u] and -.31, which are of a similar order of magnitude to that found in the medical school outcome measure. Figure 4 shows scattergrams of outcome against educational attainment separately for white and non-white students; the lowess curves are effectively parallel across almost the entire range of ability. On the 4-point outcome scale, non-white students were more likely to fail (2.7% (30/1092) vs 1.6% (48/3057), odds ratio=1.47, 95% CI = 1.108 to 1.961), and were also more likely to repeat the first year (2.9% (32/1092) vs 1.5% (45/3057)), and to pass after resits (15.6% (170/1092) vs 10.0% (306/3057)), meaning that non-white students were less likely to pass all exams at the first occasion (78.8% (860/1092) vs 86.9% (2658/3057)).

        ii) ***Male sex.*** Males performed less well on the educational attainment measure and on the overall medical school score, but better on UKCAT. Both findings have been reported previously in medical students and UKCAT candidates.

[t] The Woolf et al meta-analysis looked at results of all undergraduate examinations. Other data of our own suggest that the effects of ethnicity may be somewhat lower in the first and second years of the undergraduate course (e.g. see the Supplementary Information of Woolf et al[33]).

[u] McManus et al [39] found a significant but rather smaller ethnic effect on A-level grades in entrants to medical school.

iii) ***Age.*** Older students did less well on UKCAT but performed better at medical school, although the effect was only present for those aged 21+ and not those aged 30+. Age effects could not be assessed for educational attainment as the analysis had been restricted to those aged under 21, as there were few students older than that with data from UCAS.

iv) ***UK nationality.*** Non-UK nationals performed significantly less well than UK nationals on UKCAT, although there were no differences on the attainment measures. The reason for this is not clear, although it is possibly a language effect.

b. ***School measures.***

i) ***Selective schooling.*** There was a little evidence that students from selective schools had higher levels of educational attainment (although it must be remembered that this analysis is for those who have already entered medical school, and selection in large part would have been on educational attainment). However students from selective schools also performed significantly *less* well at medical school, but had performed *better* on UKCAT.

ii) ***DFES contextual measures of schooling.*** Although the KS5 value-added measure for schools had little relationship to the key variables, the average points gained by students at a school related positively both to individual educational attainment and individual performance on UKCAT, but related negatively to performance at medical school. Average points per exam entry showed a large effect only in relation to medical school performance.

c. ***Social measures.*** Only SEC is an individual level measure (strictly for the parents or the family), and the other measures are contextual.

i) ***Socio-economic classification.*** There was a significant effect of SEC on Educational Attainment and on UKCAT performance, but not upon medical school performance.

ii) ***Deprivation measures.*** The overall deprivation measure, as well as the seven domains, show mostly similar results to one another, for most there being an association between more deprivation (lower deciles) and lower educational attainment and UKCAT performance. In contrast, most of the indices did not relate to medical school outcome, with the particular exception of the crime and living environment measures.

d. ***UKCAT measures.*** These measures were to do with the taking of the UKCAT test, three measures being individual and one being contextual.

i) ***Numbers of questions skipped/missed, and extra time allowed.*** Hardly surprisingly, candidates who omitted to answer questions did less well on UKCAT, as also did those given extra time. Perhaps of greater interest is that neither effect manifested either in educational or medical school attainment, suggesting that the behaviours did not generalise to other situations.

ii) ***Day of taking UKCAT.*** Students who took UKCAT early in the testing cycle did better on the test[v], and interestingly that same effect also manifested in the measures of educational attainment and medical school outcome. The reasons for that will be returned to later.

iii) ***School experience with UKCAT.*** Some students come from schools in which many other candidates have taken UKCAT, and hence there is more knowledge and experience of the nature of the test. The students from schools where many candidates had taken

[v] If candidates taking the test later had performed better then question leakage may have been suspected, but that is excluded by the present relationship.

UKCAT in fact performed slightly less well, which is incompatible with any social learning process helping those at schools which have more experience, and is the opposite direction of effect to that reported by BMAT [18].

2. ***Multivariate analyses.*** Many of the 22 background variables in Table 5 are correlated with one another, making interpretation difficult. A series of multiple regressions were therefore carried out. Given that there were 22 predictor variables, and 3 dependent variables, making 66 analyses, and given also that there was a large sample size and hence quite tiny effects could be significant, and forward entry was used on an exploratory basis, a criterion of p<.001 was set. The analysis was carried out on data in which EM imputation of missing values had been carried out.
    a. **Medical school outcome.** Forward entry regression of medical school outcome on the 22 measures of table 5, found four significant predictors:
        i) ***Ethnic2.*** Non-white students performed less well than White (b= -.146, p<.001).
        ii) ***DfES measure of average points per examination.*** Beta = -.093, p<.001.
        iii) ***Percentile day of taking UKCAT.*** Beta= -.089, p<.001.
        iv) ***Student aged 21+.*** Beta =.057, p<.001.
    b. **UKCAT total score.** There were four significant predictors.
        i) ***Number of UKCAT items skipped***. Beta=-.296, p<.001.
        ii) ***Ethnic 2.*** Non-White students performed less well, beta = -.160, p<.001.
        iii) ***DfES measure of average points per student***. Beta = -.093, p<.001.
        iv) ***IMD2 employment decile.*** Beta= .064, p<.001. The higher the local rate of employment the better the student had performed at UKCAT.
    c. **Educational Attainment.** There were three significant predictors.
        i) ***DfES measure of average points per student.*** Beta = .138, p<.001.
        ii) ***Percentile day of taking UKCAT.*** Beta= -.090, p<.001.
        iii) ***Ethnic2.*** Non-white students performed less well, beta = -.134, p<.001.

Taking these analyses together, there are several points worth noting.

a. There is little doubt that **ethnicity** and **sex** need to be included in subsequent analyses.

b. The **DFES measures of average attainment at a school** would seem to be important. The two measures, average points per student and average points per entry are highly correlated (r= .665), and both also correlate highly with going to a selective school (points per student, r=.562; points per entry, r=.708). Teasing apart any underlying relationships is therefore not straightforward, and one needs to bear in mind the important and very much larger study by HESA[14,34] in which schools were divided into quartiles of achievement. The statistical analysis showed that almost all of the effect was actually secondary to going to a selective school (and the selective schools were almost entirely in the top quartile). On that basis, **Selective School** will be put into the analyses, with **points per student** also being available, to check which of the two is the better predictor. Selective School attendance has the advantage also of being much easier to interpret and explain.

c. **Age** is clearly important, but given that measures of educational attainment are not available for most students over the age of 21, the best strategy is to carry out separate analyses for students over and under 21.

d. Only one of the **IMD deprivation measures** has come through as significant in the regressions, and that is IMD2, which is a measure of employment. While this might be important, it will be omitted at the present stage in order not to complicate the analyses too much.

e. The most complicated measure is the **day of taking UKCAT**. The measure was not put in for any strong theoretical reason, but it does seem to correlate with outcome. The reasons for this are far from clear. Figure 5 shows a scattergram of UKCAT total score against date of testing. The relationship looks somewhat non-linear, being relatively flat for most of the testing season, followed by a sudden decline in the last few weeks. Many interpretations are possible, one of which would be that weaker students decide late in the day to apply for medical school, and therefore take the exam at the last minute. That would be supported by a regression showing that those taking UKCAT later have significantly lower GCSE grades (which would have been taken a year before UKCAT), and also after taking GCSE grades into account, significantly lower AS-level grades (the results of which would have come out over the summer, after the UKCAT testing season had started). Despite the strong effect, and it also being present for the two attainment measures, lack of theoretical understanding here means it is probably best omitted from the analysis at the present, the suspicion being that it is in effect a surrogate for academic ability.

Based on the analyses of the background measures the only background measures which will be included are: **Sex**, **Ethnicity**, and the **DfES** measure of average points per student at the school. Age is also important, and therefore separate analyses will be carried out for those under 21 or those over 21. An additional measure which will be included is **AlevelsHighers**, a dummy 0/1 variable indicating whether A-levels or Highers were taken during secondary schooling (in effect whether secondary education was in E&W or Scotland), as it has already been shown that Highers are better at predicting medical school outcome than are A-levels.

The relationship of UKCAT and Educational Attainment to Medical School Outcome.

After many preliminary, but necessary, analyses of a wide range of background variables (and the richness of the UKCAT database justifies such a thorough approach), it is possible to approach the key questions concerning UKCAT and its predictive ability.

Strategy of Analysis

In order to assess the predictive validity of UKCAT one needs to know several things about it:

1. ***Predictive validity of UKCAT alone, and in relation to background variables.*** Does UKCAT on its own predict outcome at medical school? If so, does this predictive validity vary in relation to student characteristics such as sex, ethnicity, age and type of schooling , and does predictive validity vary between medical schools.
2. ***Predictive validity of Educational Attainment alone, and in relation to background variables.*** This analysis is a baseline against which the incremental validity of UKCAT can be compared.
3. ***Incremental predictive validity of UKCAT in conjunction with Educational Attainment.*** This is a key analysis, since Educational Attainment has long been known to predict outcomes at medical school (and that is the basis for using it in selection), and if it is to be useful, UKCAT needs to add additional value over and above existing measures of Educational Attainment.

1. **Predictive validity of UKCAT alone, and in relation to background variables**.

a. ***Multilevel modelling.*** The basic analysis is a multilevel model in which UKCAT and the medical school outcome variables, as well as other non-categorical variables, have been standardised (i.e. mean of zero and SD of one) at the level of the medical school and the cohort, although cohort is not analysed further here. Because of the standardization there can be no variation in the intercepts of variables, which are always zero; there can however be variation in slopes. In its most general form the model is fitted at four levels:

   - Medical school country (Scotland vs E&W). [MSC]
   - Medical schools within country (MS) [MSwC] [MS/MSC i.e. Medical school within Medical School Country]
   - Country of Secondary Education (Scotland vs E&W) [CSE]
   - Individual students [Ind]

   In practice, there are almost no effects either of Medical School Country or Country of Secondary Education, and therefore they will be ignored, except when they are occasionally of significance or interest. Overall there are no significant random effects of Medical Schools within Country, and therefore instead of presenting detailed MLwiN results, summaries of fixed effects will be presented within the text.

   Variable Names. In order not to make the text too cluttered, variable names have been abbreviated and simplified as follows: OverallMark **= Outcome**; MedSchool = **MedSch**; MedSchoolScotland = **MedSchScot**; zUKCATtotal = **zUKCAT**; UCAS.Ethnic2 = **NonWhite**; UCAS.Male = **Male**; CAND.AgeGt21 = **Age21**; AlevelsHighers = **SecEdScot**; zEducationalAttainment = **zEdAttmt**; ZDFES.AvePointStudent = **zSecSchPts;** SelectiveSchool = **Selsch.**. In addition in MLwiN there are variables labeled **StudentID** (a unique label for each student), and **Cons** (the constant which is one in all cases).

   A series of nested models is fitted within *MLwiN*.

   i) ***UKCAT only.*** The initial model contained only Outcome, which was predicted by zUKCAT, (beta = .142, SE = .016), which is highly significant[w], with the beta value an estimate of the validity coefficient. Allowing the slope of zUKCAT to vary at any of the four levels did not result in any improvement in fit, and so it can be concluded that the predictive value of UKCAT is equivalent across Medical Schools, and between secondary and medical education in Scotland and elsewhere.

   ii) ***Age of medical students.*** In the current analysis measures of Educational Attainment are mostly only present for recent school-leavers. Although in this case that is due to the way in which UCAS collects its data, it is also the case in practice that medical schools are much less likely to consider attainment at secondary education when selecting medical students who are mature (in UCAS's sense of being aged over 21 at admission). UKCAT is administered to such applicants, and it is important to know whether UKCAT predicts equally well in mature and non-mature students.

      (1) The effect of age on Outcome was modelled. Including only zUKCAT and Age21 in the model, there are highly significant effects of Age21 (Age: estimate = .326, SE=.045), with no evidence of variance in the effect of age between medical schools

[w] *MLwiN* provides for all parameters an estimate and a standard error of the estimate in the form "0.142 (0.016)", where the estimate is 0.142 and its standard error is .016. A *t* (z) statistic can be calculated as *estimate/SE*, and a probability obtained. In practice it is more straightforward to treat as significantly different from zero with p<.05 those estimates for which the estimate is more than twice the size of the SE. Estimates in *MLwiN* are unstandardised, and standardised estimates can only be obtained by standardising variables prior to entering them into the program. That has been done for a number of the variables here, and when it is the case the estimate will be described as *beta*.

(or an interaction with zUKCAT). When Age21 is taken into account the beta for UKCAT increases to 0.147 (se .016).

(2) A statistical interaction between zUKCAT and Age21 was included, and was highly significant (estimate = .113, SE=.044), with no evidence of differences of slopes of any effects between medical schools. Figure 6 shows scattergrams for the raw data, and it can be seen that the slope (validity) is much higher for the mature than the non-mature students. Regression on the raw data found validity (beta) coefficients of .137 for the 4076 non-mature students, but .252 for the 689 mature students. Pearson correlations for non-mature and mature students were .125 (n=2121, SE= .021, 95% CI .083 to .167) and .234 (n=864, SE = .032, 95% CI .171 to .297) for zUKCAT (i.e. standardized within medical school), and for raw UKCAT scores were .107 (n=2121, SE = .021, 95% CI = .065 to .149) and .202 (n=854, SE=.033, 95% CI = .138 to .266).

iii) ***Influence of sex, ethnicity and schooling.*** The final model took into account the expected effects of being male, non-white and coming from a secondary school with high examination results. Effects were found for being Male (estimate= -.014, SE=.031), and being non-White (estimate= -.274, SE=.036), but there was no interaction between being male and being non-White (estimate=.091, SE=.072). On their own there were significant effects of both SelSch and zSecSchPts (estimate=-.045, SE .018),each showing significant effects, but when entered into the model together there remained a highly significant effect of SelSch (estimate = -.175 , se = .041), but no significant effect of zSecSchPts (estimate = .016, se=.023). zSecSchPts was therefore dropped from the model. No effects showed any evidence of variance in effect size between medical schools. Interaction terms were also fitted for Age21 with Male, NonWhite and SelSch, and none was significant.

iv) ***Summary.*** A number of important effects have been found:

(1) On its own, UKCAT predicts outcome in the first year at medical school, with an overall validity of 0.142 (p<.001).

(2) UKCAT on its own has a higher validity for mature students (.252) than for non-mature students (.137)[x].

(3) Other factors also influence outcome, students performing significantly better who are mature, who are female, who are White, and who have not been to selective schools. None of these factors interacts with age. Taking these factors into account does not influence the overall predictive effect of UKCAT (.138).

(4) The multi-level modeling found no evidence of differences in the predictive effects of UKCAT, Age, Sex, Ethnicity, and School type between different medical schools.

2. **Predictive validity of Educational Attainment alone, and in relation to background variables**. Adequate measures of Educational Attainment were only available for students who were non-mature, and the analysis has been restricted to these 3539 cases (with EM imputation of missing values as necessary). It was also necessary to restandardise educational achievement and outcome within cohorts and medical schools as non-mature students performed less well overall than did mature students.

a. ***Educational Achievement alone.*** Educational achievement alone, as might be expected from earlier analyses, is a highly significant predictor of first year medical school outcome

[x] We note that the UKCAT total score has a somewhat greater standard deviation in the mature entrants (237.7) than in the entrants aged less than 21 (SD = 199.9; Levene's Test, F=29.68, p<<.001), which may in part account for some of the effect.

(beta=.391, SE = .016). The preliminary analyses had suggested that measures based on Scottish Highers may be better predictors than those based on A-levels, and that was confirmed in the multilevel model, with a significant variance component in relation to place of secondary education (estimate = .027, SE=.009). No other significant differences in prediction were found in relation to medical school or country of medical school.

b. ***Sex, Ethnicity, Schooling.*** The effects of sex, ethnicity and schooling were assessed after taking into account the predictive effect of educational achievement. Overall males did less well (estimate = -.073, SE = .031), non-White students performed less well (estimate = -.201, SE = .035). Schooling was included as two measures, and both were significant, those who had attended a Selective School performed less well (estimate = -.148, SE=.029), and there was a separate, independent effect of the DfES measure of Average Student Scores (estimate = -.057, SE = .023).

c. ***Multilevel modelling.*** There was no evidence that the effects of ethnicity, school type, DfES average score (or indeed sex) showed variance at the medical school level.

3. **Incremental predictive validity of UKCAT in conjunction with Educational Attainment**. UKCAT and Educational Attainment are not independent of one another – they are correlated – and therefore a key question is whether UKCAT predicts outcome at first year of medical school after taking into account educational attainment – the incremental validity. In the present study UKCAT and Educational Attainment correlate 0.186 (n=3432, p<.001), with the correlation being slightly higher for A-levels (r=.193, n=2717, p<.001) than for Scottish Highers (r=.156, n=715, p<.001). The correlations are less than those reported by James et al[35], who reported a correlation between UKCAT total score and A-level total score of 0.392, probably because of restriction of range in both UKCAT and A-level scores in the current data. Incremental validity was assessed using the same 3,539 students as in the previous analysis, whom it will be remembered are non-mature. As with the previous analysis, UKCAT was also restandardised within medical schools and cohorts to take into account the fact that non-mature students performed somewhat better on it than mature students.

a. ***UKCAT with Educational Achievement alone.*** In view of the importance of this analysis, it is done both for Educational Achievement and also for 3 best A-levels and 5 best Highers.

i) Educational Achievement was modeled with a significant difference in slope for those taking A-levels or Highers. There was a significant effect of UKCAT (beta= .048, SE=.016), and no evidence that the effect differed according to medical school, country of medical school or country of secondary education. With UKCAT in the model, the effect of Educational Achievement was 0.412 (SE = .019) (with a variance effect due to Alevel/Highers of .025 (SE = 0.009)).

ii) Best A-levels/Highers. Mean substitution was used to allow best A-levels and Highers to be entered into the analysis as separate variables (since they are not standardized). Together A-levels and Highers were jointly significant (R=.133), although there was no independent effect of Highers, either on its own or jointly. UKCAT when entered after A-levels and Highers was significant (beta=.118, p<.001).

Taking the analyses together of Educational Achievement and best A-levels/Highers, it is clear that UKCAT has little predictive effect when the composite measures of educational achievement are included (beta= .048), but has slightly more predictive validity when only best A-levels and Highers are included (beta=.118), the difference mainly resulting from there being little variance in either of the two best measures of A-levels or Highers.

b. ***Sex, Ethnicity, Schooling.*** With UKCAT in the model, sex, ethnicity and the two schooling variables had similar effect sizes to previously (Male: -.080, SE=.031; nonwhite= -.189, SE=.035; Selective School=-.149, SE=.039; DfES measure of Average Student Scores = -.060, SE= .023). With these background factors in the model the beta for Educational Achievement was 0.419 (SE = .020), and the beta for UKCAT was .045 (SE = .016).

c. ***Multilevel modelling.*** As before, there was no evidence that the effects of ethnicity, school type, DfES average score or sex showed variance at the medical school level.

d. Summary of multilevel models of Validity and Incremental Validity of UKCAT.

i) On its own, UKCAT has a validity of about .142 overall, with a somewhat lower value of .137 in non-mature students (<21) and a notably higher value of .252 in mature students (≥ 21).

ii) Educational achievement, using the measures described in the preliminary analyses, has a substantial predictive validity, of about .391 overall, with a rather higher value for those having Scottish Highers (.453), than those having A-levels (.368). These validity coefficients apply only to non-mature students.

iii) The high validity of educational achievement means that inevitably the incremental validity of UKCAT will not be very high, and it has a value of .048 without other factors in the model, or .045 after taking account of sex, ethnicity and school type.

iv) Incremental validity could also be calculated against three best A-levels (which correlated .177 with outcome), and five best Highers (which correlated .121 with outcome). Incremental validity of UKCAT was .102 when considered against three best A-levels, and .073 when considered against five best Highers.[y]

v) Considering the background factors, there was clear evidence that males and non-White students performed less well. There were also strong independent effects of two schooling measures. After controlling for educational achievement candidates who had attended independent schools performing less well, as also did those where the DfES reported high average achievement of school students.

vi) The multilevel modelling confirmed that all of the effects described above are of similar magnitude in all of the twelve medical schools which took part in the study.

Comparison of predictive validity of UKCAT and educational attainment in the present study with previous studies.

The present estimates of the predictive validity of UKCAT, and of the educational attainment measures collected in the present study, can be compared with the weighted estimates of the correlations found in previous analyses (see table below).

	Previous medical school studies	**Current UKCAT-12 study**	University courses in general
Educational attainment score (three best A-levels/ five best Highers)	**.31**	**.39**	**.38**
Ability/aptitude scores	**.13**	**.14**	**.14**

[y] The incremental validities of three best A-levels and five best Highers, after taking UKCAT into account, were .160 and .109.

The similarity of the present study with previous ones, summarized earlier and in the appendix, is very high. The educational attainment measures calculated here have a somewhat higher correlation with outcome, .39, than do previous studies in medical schools, but the value is very similar to that for universities in general. The difference may reflect the wider range in the educational attainment scores derived here, particularly for those whose secondary schooling was in Scotland, and had somewhat less of a ceiling effect. The correlation of UKCAT with outcome, .14, is very similar indeed to the correlations found both in other medical school studies and for aptitude tests for university degrees in general. Finally, the incremental validity of UKCAT after taking educational attainment into account, is low at 0.05. Whilst we have not been able to find an estimate of the incremental validity of UKCAT over educational achievement in previous studies, the figure of .05 is similar to the figures found in other large studies concerned with the incremental validity of aptitude tests over and above educational attainment.

The predictive validity of the subscales of UKCAT and the Theory and Skills Outcome measures

UKCAT is comprised of four sub-scales. One of the original aims of UKCAT (see Introduction) was to ask whether different sub-scales might predict medical school outcomes differently, with a possibility being that some subscales worked better in some medical schools than others. The reliabilities of the subscales have been presented previously, with AR having the highest reliability (.82), followed by QR (.71), VR (.66) and DA (.58). It is worth noting, though, that the four tests have different numbers of items (60, 36, 40 and 26), so that it is not surprising that AR is the most reliable. If one uses the Spearman-Brown formula to calculate the likely reliability of a test of 162 items, the same as the current total test length, but based solely on items from each specific subtest then the expected reliabilities would be .925 (AR), .917 (QR), .887 (VR) and .896 (DA), with a mean reliability of .908 (which is a little higher than the current reliability for the total test of .860).

Table 6 shows the correlations between the sub-scales and the total score in the population of entrants. All four sub-scales correlate with the total score (not surprisingly since the total is the sum of the sub-scales). Of more interest is that the sub-scales correlate with one another to a surprisingly low degree, perhaps below what might have been expected given their internal reliabilities (described earlier) and the existence of a single, large underlying factor; the disattenuated correlations are shown in brackets. The reasons for the lowish correlations are not clear at present, The implication is that perhaps, in contrary to what might have been expected, the four sub-scales are not primarily measuring aspects of general mental ability (*g*). This raises the possibility that the sub-scales reflect real differences in respondent ability profiles, and hence there may be additional predictive information in the subscales. Table 6 also shows the correlations with EducationalAchievement, and it is striking that all four sub-scales correlate with it to very similar extents, with a slightly lower correlation for the Verbal Reasoning subscale.

So far in this report analysis of outcomes has been restricted to *OverallMark*, but there are also two other outcome variables, *TheoryMark* and *SkillsMark*. The precise components of the latter are not clear, and they were not specified to us in any more detail, but they are presumably a mixture of laboratory and communication skills, and contrast with the Theory Exams which are presumably a mixture of essays, multiple-choice exams, etc. Table 7 shows that the Theory mark correlates more highly than the Skills Mark with all of the measures, both UKCAT and Educational Attainment. The implication is that the content of the Skills exam overlaps less with previous academic measures, presumably because it is less of a standard academic assessment. Particularly striking is the difference for Verbal Reasoning, which correlates strongly with the Theory Mark, but has no significant correlation with the Skills Mark. If Skills were to include Communication Skills then that would be unexpected, since if communication were to be expected to correlate with anything it would surely be linguistic ability in a social context.

A series of multiple regressions explored the relationships in more details.

1. **Relationship of OverallMark to the UKCAT sub-scales.** Simple correlations of the four sub-scales with Overall Mark were .120, .069, .081 and .086 for Verbal Reasoning, Quantitative Reasoning, Abstract Reasoning, and Decision Analysis respectively (p<.001, N=4811 in all cases). Regression of OverallMark on TotalMark showed a strong correlation (beta = .148), but none of the subscales then contributed any additional significant variance. Regression of OverallMark on the four sub-scales alone showed that all four contributed significant variance, the largest effect being for Verbal Reasoning (beta=.093, p<.001), followed by Decision Analysis (beta=.061, p<.001), Abstract Reasoning (beta=.051, p=.001), and finally Quantitative Reasoning (beta=.037, p=.012). This ordering is not merely that of the number of items (or equivalently the reliabilities of the subscales), suggesting that it is indeed the content of the items which is important.
    a. **Multilevel modelling.** A MLM was fitted with outcome as the dependent variable, individual student and medical school as the two levels, and the four UKCAT subscales as predictors. There was no evidence of any significant variation at the level of medical schools, suggesting that the four subscales have equal predictive validity in all medical schools.
2. **Relationship of TheoryMark and SkillsMark to UKCAT sub-scales**. The theory and skills marks are correlated (r=.566), and therefore 'pure' components of each were generated for each by partialling out the other. When that was done, the only significant predictor of the Theory Mark was a higher Verbal Reasoning mark (beta=.162, p<.001), and the only significant predictor of the Skills Mark was a *lower* Verbal Reasoning mark (beta= -.075, p<.001).
3. **Effect of UKCAT subscores on OverallMark, after taking EducationalAttainment into account.** EducationalAttainment was a strong predictor of OverallMark (beta=.362, p<.001). When the four UKCAT subscores were entered into the analysis as well, Verbal Reasoning was fairly significant (beta=.047, p=.004), and Decision Analysis was just significant (beta = .036, p=.030).
4. **Effect of UKCAT subscores on TheoryMark and SkillsMark, after taking EducationalAttainment into account.** As before, the pure components of Theory Mark and Skills Mark were considered, after partialling out each from the other. Educational Attainment was then a highly significant predictor of Theory Mark (beta=.223, p<.001), but a barely significantly predictor of Skills Mark (beta= .047, p=.045), Educational Attainment predicting Skills much less strongly than it predicted Theory. When the four UKCAT subscores were then put into the model, only higher Verbal Reasoning scores predicted TheoryMark (beta=.120, p<.001), whereas only *lower* Verbal Reasoning scores predicted SkillsMark (beta= -.091, p<.001).

Summary. The four sub-scales of UKCAT do not predict the outcome measures equally. All four sub-scores contribute to predicting the Overall Outcome Score, although Verbal Reasoning and Decision Analysis had more impact than the other scores, and only Decision Analysis had an additional effect once Educational Attainment was taken into account. The Theory and Skills Marks when expressed as 'pure' marks were predicted only by Verbal Reasoning, although the Theory mark correlated with *higher* Verbal Reasoning scores, and the Skills mark with *lower* Verbal Reasoning scores, effects which persisted even when Educational Attainment was taken into account. The Skills mark had only a small association with Educational Attainment.

Distinguishing students in the four outcome groups.

Most of the present analysis has considered a continuous outcome measure. However, as described earlier, students can also be divided into four groups for whom the outcomes are quite different: those who pass all their exams first time; those who pass only after resits; those who have to repeat their first year; and those who fail and leave the medical school (and the latter group typically contains two rather disparate subsets, those with academic problems and those with motivational, health or other problems, which cannot be distinguished in the present data). There is also an argument that the most important distinction is between those who fail outright and leave the medical school, and those who do not.

Table 8 compares the four groups on the various measures used in the analysis, the continuous outcome measures (overall, theory and skills), the UKCAT score and sub-scores, and the measure of educational attainment. In most cases the measures have been standardised within schools (and it should be remembered, for instance, that entrants to different schools differ in levels of educational attainment and in their UKCAT scores). However, the raw UKCAT total score is also presented, not least because it is a number that most admissions tutors will understand since it is on the usual scale.

In general, Table 8 shows linear trends across the four categories, the better performing groups having higher overall marks, both in theory and skills exams, higher overall scores on UKCAT overall and on most of the subscores (with the exception of quantitative reasoning), and in educational attainment. A more detailed comparison suggests that often the three groups resitting exams, repeating the year or failing are not so easily distinguished, but that in part reflects lower sample sizes and hence less statistical power. There is a tendency sometimes, though, for the group who failed and left the medical school to have performed rather better than the group who repeated the first year (and that is the case in eight of the eleven measures shown in Table 8, although in no case is a post hoc comparison significant). That might be expected if the group consists of a mixture of those with very poor academic performance at medical school, and those who are academically capable but are poorly motivated. In contrast, the group repeating the first year can be assumed mostly to have adequate motivation, but perhaps insufficient academic ability.

Table 9 makes a direct comparison of the 55 students leaving for academic reasons and the 50 leaving for non-academic reasons, as classified by medical schools. Only two of the variables show significant or near-significant differences, those leaving for non-academic reasons having slightly lower course marks, but higher marks on prior educational attainment. Those leaving for non-academic reasons were also rather more variable on several of the measures.

Although interesting, and despite the large sample size in this study, with nearly 5,000 students, fewer than 100, about 2% or so, were outright failures. That number is small, and inevitably limits the power of any analysis to predict who will be in the group that fails completely. A simple t-test comparing the failures with all other students, found a highly significant difference in prior educational attainment (t(3430)=3.61, p<.001, means = -.441 vs .008) and a just significant difference on the raw UKCAT total score (t(4809)=2.022, p=.043; means = 2492 vs 2535), although not on the standardised UKCAT total score. The effect of the UKCAT raw score was not significant after educational attainment was taken into account. There were of course also very large differences on the overall mark, theory mark and skills mark at medical school (-2.64, -1.26 and -1.08 vs .031, .098 and .084 respectively, all p<<.001).

As a guide to the predictive ability of UKCAT and A-level grades, figures 7 and 8 show the results of logistic regressions looking at the probability of a student passing all exams without resits, passing the first year satisfactorily (with or without resits), having major problems (either having to retake the first year or having to leave the medical school), or being required to leave the school, in relation to three best A-level grades and total UKCAT score. In both figures the range on the abscissa has purposely been set to the entire possible range of marks (EEE to AAA for A-levels or 900 to 3600 for UKCAT) and prediction is outside of the actual range of marks (which are shown by grey bars at the top of the figures. Within the range of applicants it can be seen that A-levels have a much greater predictive value than UKCAT. The latter is particularly important if entrants are being considered with non-conventional qualifications, such as those with typical CCC grades in the EMDP at King's [36]. Although claims have been made that performance in these students is indistinguishable from standard entrants [36], more detailed study suggests that their average performance is about .73 standard deviations below that of contemporaries on the conventional five-year programme [37,38]. That figure is compatible with the lower expected performance suggested by figure 7.

Corrections for restriction of range and attenuation due to low reliability

Although the *calculation* of a correlation between a **selection measure** (such as educational attainment or a UKCAT score) and a **performance measure** (such as first year achievement in medical school) is straightforward, the *interpretation* of such a correlation is far from straightforward.

1. ***Description and discussion of approaches to range restriction.***

    a. **Restriction of range.** The primary problem, known as **restriction of range**, which is particularly difficult to handle, is that selection occurs across *the entire range of* **applicants**, whereas the correlation between the selection measure and the performance measure can only be calculated *in the restricted group of individuals who are selected, the* ***entrants,*** or as they are sometimes called, the **incumbents**. Calculation of correlations and other measures are typically referred to as being in the **restricted population** (the entrants, the incumbents, those in medical school), or the **unrestricted population** (the group of applicants for the course). A subtle but important point is that the pool of applicants is not a fixed group, but as well as concerning the actual applicants in any particular cohort, may also be extended to consider the pool of possible applicants. As an example, the majority of entrants (and hence applicants) for UK medical schools in recent years have had high A-level grades, typically ABB or above. However a recent move in some schools, as a part of widening participation programmes, such as the Extended Medical Degree Programme at King's, has allowed entrants who typically have A-level grades of CCC [36]. The perception of such entrants by future candidates may well mean that future applicants will apply with much lower A-level grades than at present, and calculations need to be made as to their likely performance on a medical course. In the extreme case, it might be argued that widening participation should allow any applicant from the general population, irrespective of their overall ability, and calculations would then need to be made of their likely performance in the first year at medical school.

    b. **Attenuation and reliability.** A secondary problem is that all measures, be they the selection measure or the outcome measure, are imperfect, being *unreliable* to a greater or less extent. A **reliability coefficient** reflects the proportion of **variance due to the true score** in relation to the **total variance** comprising the true score and **random, measurement error**. A coefficient of 1 indicates perfect reliability, with no measurement error, and a coefficient of 0 indicates a complete lack of reliability, where the measured scores are no more useful than random numbers.

Reliabilities are not easy to interpret (see Cortina [39] and Schmitt [40] for some problems of use and interpretation). One of the problems is that reliability and restriction of range are not independent. Restriction of range reduces reliability, because it reduces true score variance but leaves measurement error unaffected [z]. As a result, a measure that is highly reliable in an entire population is often quite unreliable in a subset of the population with reduced range.

The performance measure typically also has the special problem that it is *only* measured in those who are selected (and for instance, there is little indication of how, say, current candidates with EEE at A-level would perform in first year medical school exams because such candidates are never selected).

c. **Disattenuation.** Correlations can be corrected straightforwardly for lack of reliability, where the process is known as **disattenuation**. As an example, if one test has a reliability of .6, and a second test has a reliability of .7, and the empirically measured correlation between the two tests is.5, then the **true-score (disattenuated) correlation** is .5/√(.6 x .7) = .772. Although it might be tempting to say that one measure accounts for only $.5^2$ = 25% of the variance in the other, the disattenuated correlation shows that in principle, one variable accounts for $.772^2$ = 59.6% of the variance in the other. If better, more reliable, tests could be used then in principle up to six-tenths of the variance could be accounted for, rather than one quarter. The true-score correlation also shows that 100-59.6 = 40.4% of the variance in one measure cannot be accounted for by the other test, and that as a result *some second process must also contribute to the relationship between the two tests*. In that case, although it might be worth investing effort in improving the reliability of the tests, it would also be sensible to look for other, independent, processes that contribute to the relationship. Disattenuated correlations, also known as **construct-level correlations**, therefore give a proper interpretation of the strengths and weaknesses of relationships after taking account of the effects of measurement error.

d. **Direct and indirect restriction of range.** Although simple corrections for reliability are straightforward, correcting for reliability in presence of restriction of range is far harder, and the formulae are very non-linear. Although there is a large literature on this, going back to Karl Pearson in 1903, and then to the influential work of Thorndike [41], it is fair to say that much of it is incorrect and misleading, as Hunter et al [42] (see also Le and Schmidt [43]) have emphasised. The problem is due to a critically important distinction between **direct restriction of range** and **indirect restriction of range** (for a review see Ghiselli et al [44], with there also being a problem with the order in which the various corrections are applied [45]). To put the practical problem simply, as Hunter et al say, "range restriction is indirect in almost all validity studies in employment and education" (p.600). Because the equations, as proposed by Thorndike, were only soluble in practice for direct restriction of range, most researchers used the formulae for direct restriction of range instead. As Hunter *et al.* say, "although it was known that [the direct restriction formulae] undercorrected, it was apparently presumed that the undercorrection was modest (e.g. 5% or so). As it turns out, it is not" (p.600). Hunter et al [42] provided practical solutions to the problem of indirect restriction of range, and it is their formulae,

[z] A thought-experiment shows the problem. Consider a measure with a high reliability in the population, from which a group is selected in which every individual has the same true score. Measurement error still applies, and hence there will be variance in the group members' obtained scores. But since the true scores are all identical, all of this variance in the obtained scores is error. Consequencely the reliability of the scores in the group selected is zero, despite the reliability being high in the population as a whole.

which do not require information about the indirect process of selection, that will be used here.

e. **Construct-level validity and operational validity.** If measures are separated in time (e.g. at selection and during the first year of a course) then the relationship between the scores is not merely a correlation but can also be regarded as a **validity coefficient**, since **predictive validity** is conceptualised in terms of the correlation between measures at time 1 and subsequent measures at time 2. Although **true-score or construct-level correlations/validities** are useful scientifically for assessing the true relationships between measures, it is useful also to consider what is called the **operational validity** of a selection measure, which allows for measurement error in the selection measure but not in the performance measure, on the basis that selectors have control over selection measures but may not have control over performance measures, and hence the operational validity gives a sense of the practical utility of a selection measure.

f. **Incremental validity.** Often in selection there are several measures which are being considered, and an important practical issue concerns the incremental validity of measures after taking one or more other measures into account. Incremental validity describes the extent to which a measure has a unique contribution to make, after other measures have been considered. Incremental validity, like other validity coefficients, is affected both by lack of reliability and restriction of range. Without giving details, Hunter et al [42] sketch out straightforward ways in which incremental validity can be assessed, using a multiple regression approach using correlations disattenuated for reliability and range restriction.

2. *Calculating validity coefficients corrected for reliability and restriction of range*.

In order to carry out the corrections for range restriction and reliability, so that the construct-level validity can be calculated, the equations require a number of input parameters, the details of which need specifying. Table 10 provides a summary of the various values.

i) **Estimation of parameters for three best A-levels.** Validity coefficients will be calculated only for the standard measure of three best A-levels, since the necessary statistics for the population are available for these measures and not for Scottish Highers or for the more sophisticated Educational Attainment scores also reported here. In order to apply the corrections of Hunter et al [42] , several different values are required, and they are described below, as well as in table 10. Analysis is carried out separately for two situations, when the applicant pool is those who actually applied, and the potentially wider applicant pool of all UCAS applicants.

(1) ***Correlation of selection measure and medical school performance in the restricted group of medical school entrants*.** As described previously, table 7, the correlation between first year medical school performance and three best A-levels is .177. This correlation is inevitably low because of the large restriction of range on A-levels, coupled with the ceiling effect, although it is statistically very significant (p<.001).

(2) ***Range restriction on selection measure (u_X = SD(entrants)/SD(applicants)).*** The standard deviation (SD) of three best A-level grades amongst medical school entrants is 1.309. Using the data described by McManus et al [46], which looks at an entire cohort of applicants to UCAS, the SD of the three best A-level grades with the

unrestricted group being those applying to medical school is 3.904, meaning that u_X = 1.309/3.904 = .335. For a wider consideration of all applicants to UCAS[46], where the SD of the three best A-levels is about 5.99, u_X = 1.309/5.99= .219.

(3) ***Reliability of selection measure in the unrestricted applicant population.*** The selection measure is A-levels, and obtaining a reliability is not entirely straightforward for exams such as this which are modular[47]. Recent work has assessed the reliability of A-levels [48] [49,50], but estimating a reliability for individual A-level subjects is far from straightforward. Bramley and Dhawan[50] have estimated that the typical (median) reliability of an A-level component is .83 (p.96), based on the entire pool of A-level candidates, and all A-level subjects. However it is not straightforward to go from there to an estimate of the reliability of an individual A-level or indeed a cluster of three best A-levels[aa]. Analysis of AS-level qualifications [49] suggests that science subjects maybe somewhat more reliable than other subjects. For the present analysis we have therefore taken a more direct approach to calculating the reliability of the three best A-levels in the restricted population of medical school entrants. The three highest A-levels were identified, and then randomly re-ordered within students, so that the highest, middle and lowest could be arbitrarily described as A-level 1, 2 and 3[bb]. Coefficient alpha was then calculated for that restricted group and found to be .359. Although the A-level results for medical students were highly skewed, there is evidence that with large samples, alpha is little affected by skewing or kurtosis [51]. Coefficient alpha for the unrestricted groups of all medical school applicants or all UCAS applicants could then be calculated using the formulae of Hunter at al[42] (see their table 2, p.603), and gave estimates of the reliability of .928 and .969.

(4) ***Reliability of performance measure in the restricted group of medical school entrants.*** We know of no published calculations of the reliability of examination results in the first year at medical school. However the meta-analysis of Bacon and Bean [19] estimates that reliability of first-year university exams is of the order of 0.84, and we will use that estimate as the main measure here. However later we will also carry out a sensitivity analysis of the effects of that estimate being somewhat different.

ii) **Estimation of parameters for UKCAT.** The main analysis for UKCAT is for the applicant pool who actually applied. No attempt is made to estimate validity in the more general pool of applicants to UKCAT since information is not available.

(1) ***Correlation of selection measure and medical school performance in the restricted group of medical school entrants.*** As described in table 7, the correlation between first year medical school performance and total UKCAT score is .148. Unlike the correlation of performance with three best A-levels, this correlation is not unduly reduced due to any ceiling effect; it is statistically very significant (p<.001).

(2) ***Range restriction on selection measure (u_X = SD(entrants)/SD(applicants)).*** The UKCAT Annual Reports for 2006, 2007 and 2008 report SDs amongst all medical school applicants taking UKCAT of 259, 255 and 268 (mean=260.7). The standard deviation (SD) of UKCAT total score amongst the medical school entrants in the UKCAT-12 study was 206.3, so that u_X = 206.3/260.7 = 0.791. It should be noted that this selection ratio is very much less than for A-levels, reflecting the fact that

[aa] Tom Bramley, personal communication, February 2012.

[bb] This was necessary as otherwise, because the A-levels had been sorted into order, so that the three highest could be obtained, the correlations between 1 and 2 were different on average than between 2 and 3, and so on.

selection in some schools was not directly upon UKCAT scores, whereas most schools select strongly on educational achievement.

(3) ***Reliability of selection measure in the unrestricted applicant population.*** The UKCAT Annual Reports for 2007 and 2008 report reliability coefficients for the total score of UKCAT of .87 and .86 (mean = .865).

(4) ***Reliability of performance measure in the restricted group of medical school entrants.*** As for A-levels, in the absence of any better estimate of the reliability of first year examinations, we have used the estimate from the meta-analysis of Bacon and Bean [19] of 0.84, and later we will present a sensitivity analysis.

2) ***Construct-level validity for three best A-levels.*** Although the simple correlation between three best A-level grades and first year performance is low at .177, the range is very restricted due to selection in large part being on educational achievement, which also means that the reliability of A-levels within the restricted group that is selected is also reduced, so that when the low reliability is taken into account, along with the range restriction, the construct level validity is .822, which is high (see table 10). The implication is that school-level educational achievement is in fact a very good predictor of first-year medical school performance. If a wider range of A-levels were to be considered on the basis that applicants might in principle be accepted across the entire range of university applicants, the construct level validity is higher still, at .854 (see table 10). Operational validity is also good, at .822 for medical school applicants (or .916 across the entire university applicant range). The high construct level validities (.854/.930) implies that about 73% or 87% of the entire construct-level variance is explained by A-levels, leaving perhaps 27% or 13% of variance in performance in the first year at medical school to be explained by other factors. The nature of that remaining 27% (13%) is still not clear, although it must be emphasized that since the correlations are at the construct level, it is not measurement error or noise. The recent meta-analysis of Richardson et al [52] provides a number of personality and study-related measures which have been found to relate to outcomes in higher education.

3) ***Construct-level validity for UKCAT.*** The construct level correlations for UKCAT are very different from those for three best A-levels. Table 10 shows that although the simple correlations of UKCAT and three best A-levels with first year performance are broadly in the same range (.148 and .177), the interpretations of those correlations are very different. There is relatively little range restriction on UKCAT scores (u_x = .791 compared with .335 for A-levels), and as a result the reliability of UKCAT in the restricted group is high at .784. Taking the reliability and range restriction into account, the construct level validity of UKCAT is only .239, suggesting that while on its own it might predict some of performance in first year at medical school (at best, with perfect measures, about 5.7% of the total variance), the remaining 94.3% of true variance in medical school performance would be unaccounted for).

4) ***Sensitivity analysis of the reliability of first-year medical school exams.*** It was mentioned earlier that there are no very good estimates of the reliability of first-year medical school examinations, the medical schools taking part in UKCAT-12 did not provide estimates of the reliabilities of their exams. In the absence of any other estimates, the analysis of table 12 therefore uses the meta-analytic estimate of 0.84 provided by Bacon and Bean [19]. However, it may be that medical school examinations are actually somewhat more or less reliable than 0.84, and we have estimated the impact of different values of the reliability. In so doing we were motivated by knowledge we had acquired about two separate sets of first-year examination results at UK universities. One assessment consisted of first year exams in a non-medical subject, marked by conventional essays, where the reliability based on six modules was about .74 (and which, incidentally, would result in a reliability of about .90 across

a three-year course). The other estimate is from a UK medical school, with four first-year modules, mostly machine-marked, where the reliability was 0.93. The table below therefore repeats the key parts of calculation of the construct-level correlations of Table 10, but using values of .74, .84 and .93, as well as a 'perfect' reliability of 1.

Predictor	Three best A-levels	Three best A-levels	UKCAT total score
Population	Medical school applicants	All UCAS applicants	Medical School applicants
r_{YYi} = .74	.870	.939	.254
r_{YYi}= .84	.854	.930	.239
r_{YYi} = .93	.840	.923	.228
r_{YYi} = 1	.830	.917	.220

As the table makes clear, the construct-level correlations are not particularly sensitive to the particular value of the reliability of first-year medical examinations, and the pattern remains the same for A-levels and UKCAT whichever value is chosen.

5) *Corrected incremental validity of UKCAT and three best A-levels*

Hunter et al [42] comment (p.606) that the calculation of the construct-level incremental validities of measures is straightforward, using the construct-level validities for the individual variables, and the disattenuated correlation at the applicant level. The construct-level validities themselves are given in table 10. In the present data the raw correlation of UKCAT with three best A-levels is .088, but that is not the correlation that is needed, as it is in the restricted group. The only study of which we are aware which gives a correlation between A-levels and UKCAT scores is that of James et al [35] , which reports a correlation of .392, although the measure of A-levels is actually "A-level total tariff scores" (their Web Table A), in which those with four or more A-levels get higher numbers of points. If that correlation is used in the regression (and after disattenuation for unreliability it becomes 0.44), then the incremental validity for A-levels over UKCAT is .927, and for UKCAT over A levels is -.167. Those values are a little problematic, and suggest that there is multicollinearity and/or suppressor variables in the relationship. Taken at face value they appear to mean that UKCAT has a *negative* predictive value once A-levels are taken into account (i.e. at any particular A-level grade, applicants with high UKCAT scores perform *less* well than those with low UKCAT scores); although that is technically possible, it does not seem to make much sense in the present context (and is also incompatible with the small but positive incremental validity found in the uncorrected data from the entrants). A possibility is that the correlation of .392 is artefactually high[cc], and if it were 0.28 then UKCAT would have an incremental validity of zero, and for values of less than 0.28, UKCAT would have a positive incremental validity[dd]. The correlation of UKCAT and three best A-levels in entrants is .099, and therefore it seems likely that the correlation in applicants is somewhere between .099 and .392. An average of the two, of .245 (disattenuated . 273),

[cc] In the James et al paper [35] the analysis was of A-level tariff scores, with the potential problem that candidates taking four or more A-levels had much higher scores, as is shown in figure 1 of the unpublished report of Yates [53]. However it seems unlikely that including four A-levels as such has had a major impact on the correlation, since Yates also reports (her table 9) a correlation of averaged A-level tariff score (based on all exams taken) of 0.390, which is almost identical to the published correlation. The variance of the three best A-levels will be lower than the variance of all A-level grades, although the effect is perhaps not a large one, because the majority of candidates take only three A-levels. Nevertheless it would be worth knowing the correlation of three best A-levels with UKCAT score.

[dd] Were the correlation to be at its lowest likely possible level of zero then the incremental validity of UKCAT would be .239 (i.e. the same as its simple construct level validity), although that does not seem plausible.

gives the incremental validity of UKCAT to be .006. For the present, that is the best estimate that we have[ee].

6) ***Construct-level validities and incremental validities: Discussion and conclusions***. Performance in the first year at medical school is strongly predicted by prior educational attainment, once restriction of range and the consequent lack of reliability are taken into account using Hunter et al's corrections for indirect range restriction[ff], to the extent that there is little need to invoke any other factors to account for the relationship. The operational validity is somewhat less, mainly because of restriction of range in A-levels, mainly due to ceiling effects. That problem may be solved in part once A* results at A-level are available (although given grade inflation it is not clear how long they or even A** grades would continue to be effective)[54]. The picture for UKCAT is very different. The construct-level validity of UKCAT is low, at .239, suggesting that UKCAT alone is not a satisfactory predictor of medical school performance, even when lack of reliability and restriction of range have been taken into account. The incremental validity of UKCAT over A-levels is difficult to calculate, due to the absence of an accurate estimate of the correlation of A-level grades and UKCAT in the applicant population, but it is likely to be low, and may be close to zero.

Overall discussion and Summary

The UKCAT-12 study is one of the largest and most systematic in the history of UK medical education, studying nearly 5,000 entrants to twelve different UK medical schools, all of whom have taken UKCAT, and in addition have information on first year examination performance, as well as a range of background information, particularly including **educational attainment** in the form of A-levels, AS-levels and GCSEs, or Scottish Highers and Advanced Highers, **demographic factors**, and **socio-economic and school-level contextual factors**.

1. ***Reservations, limitations and constraints***. In considering these results it should be remembered that:

   a. Information on educational attainment is not available for mature students because their exams were taken before UCAS systematically startedcollecting this data. However most of these individuals must have taken A-levels or Highers, and therefore all of the results are presumably sitting in paper or electronic files within medical schools. It would be tedious but far from impossible to extract these data were it decided they were of sufficient import.

   b. Information is only available about performance at the first year of the undergraduate medical course, a year in which for most medical schools a majority of the course will consist of basic medical sciences. It may well be that patterns of prediction shift as students begin to study clinical subjects. Having said that, the general pattern of medical school performance is that performance early in a course predicts outcome in later parts of the course, despite the format and content changing. In particular the twenty-year follow-up of students from the Westminster Medical School[13], in whom both A-levels and

[ee] We contacted Professor James and Dr Yates in the hope that they can provide a more accurate estimate of the correlation from their more extensive data. However under the terms of their contract with UKCAT, all data have been returned to UKCAT, the researchers no longer have access to the data, and it is not clear how access could readily be obtained. The issue must therefore rest there for the present.

[ff] We are aware that many researchers still use direct range restriction. For completeness we note that using the Hunter et al corrections for direct restriction, the construct validities in the applicants would be .518 for three best A-levels and .215 for UKCAT. Although the absolute values differ for direct range restriction, the ultimate conclusion is that there is little variance associated with UKCAT. That is still the case when incremental validities are calculated (and the correlation of .392(0.44 when disattenuated) gives fewer problems with that calculation),the incremental validity being .525 for A-levels and -.016 for UKCAT total score.

intelligence test results were known, found that A-levels predicted career outcomes through to attaining consultant/principal posts, whereas general cognitive ability, had no predictive value in post-graduate careers. Similarly, the longitudinal data of Woolf [55] show strong correlations between performance in the first, second and third years of a medical course (pages 120-1).

c. Because of the nature of UK medical education there is no way of assessing the comparability of absolute scores at the different medical schools. It can be presumed, as in all schools, that the ordering is reliable and consistent, but there is no way of knowing whether high or low achieving students in one school are comparable to high or low achieving students in another school. That is somewhat problematic because there is little doubt that the average scores on UKCAT and educational attainment undoubtedly differ between the twelve schools. All conclusions can only therefore be taken in relation to *relative performance within schools*.

d. The authors have not been informed of which particular medical schools are taking part in the UKCAT-12 study, and in particular they have no information on which particular subjects and topics are classified as 'Theory' or 'Skills'. Neither is there any information on the type of teaching at a particular school (e.g. problem-based learning, traditional, or whatever).

e. The study has looked entirely at data from *entrants* to medical schools. We have not analysed data from *applicants*, but have attempted to correct the analyses for the 'restriction of range' which inevitably occurs because those selected are less variable than those applying, particularly on measures of educational attainment. The corrections have required using external statistics on UKCAT and A-levels, which have been incorporated into the analysis.

f. All of the measures described here are ‘unreliable’ in the statistical sense, as is the case for almost all measures in biological and social sciences. We have provided both raw correlations uncorrected for unreliability (attenuation), and also estimates of construct-level correlations, corrected for attenuation and range restriction.

2. ***Implications for medical education in general***. Because of its size and the fact that is has twelve medical schools taking part, the UKCAT-12 study, irrespective of what it finds in relation to the predictive validity of UKCAT, also provides a detailed picture of medical student performance, providing insight into a range of issues of relevance to medical education and policy.

    a. **Differences between medical schools**. Of particular interest and because twelve medical schools have taken part, multilevel modelling can be used to assess whether there are differences in the influence of specific factors in particular medical schools. If there are differences between medical schools then that presumably relates to local teaching practices, the culture of an institution, or whatever, whereas the absence of differences between medical schools suggests a more global mechanism which cannot be laid at the door of a particular institution. In practice, it can be emphasised at this point that *none of the factors showed any evidence of significant differences between medical schools*. Although much lipservice is paid to different medical schools having different cultures, etc, the current study provides no evidence that such cultural or other variation has a significant effect on the prediction of first year medical school performance.

    b. **Prior educational attainment**. There is little doubt that prior educational attainment at school predicts performance at medical school. In general that is not surprising (it is a commonplace in education that the best predictor of future behaviour is past behaviour, and the best predictor of future exam performance is past exam performance). Neither should it be concerning since doctors continue to take exams through much of their careers, and that is an important mechanism for keeping up to date. However, this is somewhat surprising in the present study because there is a general perception that there can be little predictive value in A-levels and Highers as the vast majority of students are at ceiling, with AAA or AAAAA respectively. Although it is true that the the restricted subgroup of people who enter medical school have similar A level results, it must be remembered that the key correlation is not that within this restricted group *but in the unrestricted group who apply to medical school*. Considering just entrants, our detailed analysis suggests that there is more information available in academic achievement than a simple scoring of the best three or five grades would suggest (and presumably there is more information still within the unrestricted group). In particular:

i) **A-levels, AS-levels and GCSEs**. For those taking A-levels, the grades obtained at particular subjects are predictive, and poor achievement at any A-level science seems to be predictive of a less good outcome. Performance at A-level General Studies, which is not usually counted towards A-levels, is also predictive of outcome. AS-levels, and also GCSEs also provide additional information, so that the final measure of Educational Attainment predicts first year outcome with a correlation of about .391, which is typical of many studies of university performance. In contrast, 'best three A-levels' has little variance and a poor predictive ability within the restricted group of entrants.

ii) **SQA Highers and Advanced Highers**. For those taking Scottish Highers, there was additional information present in the more detailed scoring (e.g. A1, A2, B3, B4, etc, rather than A, B, etc), and there was also further information present in Advanced Highers results. When taken together the net result was that a composite score showed a substantially higher correlation with medical school outcome than was the case for A-levels.

iii) **'Good-enough' vs 'More-is-better'**. A difficult issue in all higher education concerns the extent to which, particularly with educational attainment, ever-higher exam results result in ever-higher achievement. Few would doubt that very low educational attainment is probably incompatible with training as a doctor (EEE at A-level? A couple of low GCSEs? No GCSEs?). Above such levels there is, though, an argument that beyond a certain level of achievement (perhaps CCC or BBB at A-level), there is no benefit from additional attainment; some particular level is 'good enough' to be successful. A clear prediction from is that above some minimum level there should be no correlation of outcome with prior attainment (the curve would be concave downwards). In fact these data, as with other data elsewhere[32], make clear that higher prior attainment correlates with higher outcome across the entire level of performance (and indeed the curve is *concave upwards*, suggesting that very high prior attainment is disproportionately related to higher outcome). The rejection of the 'good-enough' model has both practical and philosophical implications for medical education.

c. **Individual predictors of outcome.** Individual characteristics of students also predicted outcome, even after taking educational attainment into account. In particular:

i) **Sex.** Male students performed less well than female students. That finding is not particularly unusual in higher education (or in secondary education), although the mechanism is not yet fully understood.

ii) **Ethnicity**. Students from ethnic minorities performed less well than did White students. The result is similar in size to that estimated in the meta-analysis of Woolf et al[29]. The mechanism of this effect is far from clear, but detailed analysis suggests that the effect is present across the entire range of educational achievement. It is also important that the multi-level modeling shows the effect to be equivalent at all twelve medical schools.

iii) **Selective schooling**. Students who have been educated at selective schools underperform relative to those who have been educated in non-selective schools, an effect similar to that reported for universities in general[34]. The mechanism is not at all clear, but possibilities include that intensive coaching at a selective school, perhaps as a result of 'hot-housing', produces a 'halo' effect so that A-level grades are somewhat higher, which may be due to learning presentational skills or whatever, so that at university, and without specific coaching, performance drops back somewhat. Alternatively, it could be that non-selective schools are less able to inspire all of their students to the same extent as are selective schools, but the university environment does have that effect, and those from non-selective schools then reach the level of those from selective schools. We have looked in detail at our results and are unable to come to any clear distinction between these possibilities. The finding that, within non-selective school medical students, there are some small effects related to value-added and mean achievement at secondary school, suggests the difference between selective and non-selective schools may be more complex than merely being a function of fee-paying or selection, but any further clarification is currently not possible.

iv) **Non-predictive factors**. Negative results are also important. Without going into the results in detail, in general there are many negative results of interest, such as the absence of effects of socio-economic group or contextual socio-economic factors on performance at medical school, once educational attainment and other factors are taken into account. That is reassuring, in that it implies that medical schools are treating all students equally, irrespective of background. It should be remembered in interpreting such results, and they have implications for how contextual factors in general are used in student selection, that these background factors are related to educational attainment itself, as well as to UKCAT scores, which are used in selection. The nature of the causal relations is probably complex.

3. ***Predictive validity of the total UKCAT score***. The primary purpose of the study was to assess the predictive validity of UKCAT and its subtests, although that needed to be done within the broader context of other measures such as educational achievement, demographic and social background, etc..

i) Performance at UKCAT overall does correlate with medical school outcome. However the correlation is not very high (r=.142 across all medical students), and that value is fairly typical of the predictive validity of aptitude tests in general, particularly aptitude tests which make no attempt to assess science knowledge or achievement.

ii) The predictive validity of UKCAT on its own is substantially higher for mature entrants to medical school (r=.252), than it is for the typical eighteen-year old applicant (r=.137). Since mature and graduate entrants can present difficult decisions for admissions tutors, because past educational attainment is not necessarily a guide to current potential in such students, UKCAT may have more utility with graduate entrants. Having said that .252 is at marginal levels for a practical test for selection[gg].

iii) Considering only non-mature students, once educational attainment is taken into account the incremental validity of UKCAT is only .048, although this value is still significantly different from zero. The practical utility of such a correlation is low, but it does probably provide important theoretical insights into performance, suggesting that

[gg] It is also not straightforward to correct the value of restriction of range or reliability.

mere educational attainment is not enough on its own to allow adequate prediction of medical school performance. Whether a more extensive, differently structured test might allow some of that variance to be made more predictive is far from clear.

iv) Students who drop out of medical school or fail overall are a particularly important subgroup. They are relatively rare, only about 2% overall having dropped out by the end of the first year in this large sample. It is noteworthy, particularly in comparison with the 2% who have to repeat their first year (and presumably are motivated but are less academically able), that those who drop out have somewhat higher educational attainment and UKCAT scores. The group of dropouts is particularly hard to identify, and it is not clear that educational attainment or UKCAT scores provide any obvious assistance. Our analyses comparing students who dropped out due to academic or non-academic reasons provides little evidence of any clear distinction between the groups.

4. ***Predictive validity of the UKCAT sub-scores***. The four sub-scores of UKCAT are correlated at surprisingly low levels, even after taking reliability into account, suggesting that perhaps they are indeed measuring separate cognitive abilities (despite the substantial evidence that sub-components of general cognitive ability, including verbal, numerical, and abstract reasoning ability, are strongly associated with *g*). Courses within medical schools require different aptitudes and abilities, and in the present study the courses were divided into Theory and Skills. The most important difference between the sub-scores was that Verbal Reasoning was the best predictor of overall outcome and of Theory exams. Interestingly Verbal Reasoning was also significantly correlated with Skills exams, but *negatively* (so those with the highest verbal scores did least well on skills exams, other factors being taken into account). We have not been informed of the nature of the various Skills exams, but if, say, they consistedof pathology or histology, requiring recognition of microscope slides or whatever, then the correlation would not be unexpected, but if the Skills exams were mainly communication and similar skills, then the negative correlation with verbal performance would be unexpected and surprising.

5. ***Comparison with other studies***. Elsewhere we will present the results of a meta-analysis of tests of aptitude (general cognitive ability and attainment) for prediction of outcome at medical school. Suffice it here to say that the predictive validity of educational attainment and of the UKCAT aptitude test is similar to that found in many other studies, as also is the incremental validity of UKCAT. That UKCAT on its own predicts substantially better in mature than non-mature students is, as far as we can tell, a novel finding, other studies not having compared the two groups on an identical test. However, in general GAMSAT, a test designed for graduate-entry courses, does seem to predict somewhat better than tests such as UMAT, which are designed for non-mature, traditional entrants. Whether the GAMSAT/UMAT difference relates to test content[hh] or to testing applicants of different ages is not clear at present.

6. ***Summary.*** The UKCAT-12 study provides a superb platform both for evaluating UKCAT itself, but also for assessing a wide range of potentially predictive factors of medical school outcome, having both high statistical power and the potential for comparing across medical schools. UKCAT itself does have some predictive validity, particularly in mature applicants, but that validity is substantially less than for prior educational attainment, and the incremental validity is low (although statistical significant).

[hh] GAMSAT has more specific science content than does UMAT.

Acknowledgments.

We are very grateful to all those working on the UKCAT project for their help at all stages of this project, in particular Rachel Greatrix and John Kernthaler, who have provided much assistance. We also thank Dr Owen Bodger for kindly providing additional correlations from the Swansea study, and Professor David James and Dr. Janet Yates for their help with the analysis of A-levels.

Appendix: Review of other studies of aptitude and attainment tests in predicting performance in medical school and at university in general

In this appendix we overview other studies of which we are aware in which the predictive validity has been evaluated for aptitude tests and attainment test. We begin by considering the predictive validity of attainment tests, firstly in medical schools, and then in higher education in general, and then we look at the predictive validity of aptitude tests, firstly in medical students and then in higher education overall.

1. Predictive value of educational attainment: Medical school studies

a. ***The meta-analysis of Ferguson et al.*** In their 2002 meta-analysis[56] of 62 published papers, based on 21,905 participants, there was an average (raw) correlation of 0.30 (confidence interval .27 to .33) between previous academic performance and undergraduate medical school outcome. The correlation was .36 if corrected for unreliability of measures and .48 if corrected for restriction of range. The remainder of the analysis will be restricted to studies not included in the Ferguson *et al* meta-analysis.

b. ***A Dutch study in a group selected by lottery.*** A recent study in Holland [57] looked at a medical school where a lottery had been used for selection (and therefore instead of restriction of range rather there was actually hyperdispersion). There was a raw correlation between educational attainment and medical school performance in 398 students of.36. This figure increased to .41 when corrected for unreliability, and .37 when corrected for the increased range in the group.

c. ***The ISPIUA study***. This study[58,59], carried out in the late 1960s, is described further below since it included an aptitude test as well. Here it is sufficient to say that in a sample of 294 medical students the correlation between first year university performance and mean A-level grade was 0.35.

d. ***The Cambridge Study.*** The Cambridge Study, also mentioned below[30] as it assessed BMAT, included 988 medical students taking Part 1A and 893 students taking Part 1B Medical and Veterinary Sciences Tripos. Correlations of outcome with best three AS-levels were 0.38 and 0.37 (median=.375) and with A* GCSEs were.27 and 0.25 (median=.26).

e. ***Arulampulam et al.*** This study[60] looked at USR (Universities Statistical Record) data for 7789 medical students entering medical school in 1985 or 1986, and who had completed or left their course by 1993. The outcome measure was dropout (10.7% of students), and a complex logistic model was fitted. A-levels (best three marks) were lower in those who did not drop out (mean = 26.3 SD 3.9) than in those who did drop out (mean=25.4, SD 4.3), giving an effect size of 0.23, equivalent to a correlation of 0.12[ii].

f. ***University of Western Australia***. In a study at the University of Western Australia[61] of cohorts of students entering from 1999 to 2009 (N=1174), educational attainment (TER) correlated .468, .401, .321, .230, .208 and .206 (median = .27) with grades in years 1 to 6 respectively.

g. ***University of Queensland***. A study at the University of Queensland[62] (N=706) found correlations of year 1 and year 4 outcomes with grade-point average (GPA) (r_s = .50 and .35).

[ii] Using the method at http://www.coe.tamu.edu/~bthompson/apaeffec.htm

h. ***The Coates study.*** In a complex study, Coates[63] looked at six unnamed Australian institutions, studying a total of 351 students. Year 1 grades correlated significantly with previous GPAs (r=.47).

i. ***The 91 cohort study***. In the 1991 cohort study of one of us[64], the data of which were analysed for the 2005 paper on aptitude tests[13], and are provided in the online supplementary information, A-level results correlated .202 with Basic Medical Sciences results (n=3112, p<.001)[jj].

j. ***The Swansea Study***. A recent study[65] (N=105) looked at GAMSAT in graduate entrants to the University of Wales. The authors kindly re-analysed data for us, and there were correlations of .331 and .368 (n=97 and 100, p=.001 and <.001) of GCSE and A-level results with first year examination marks.

k. ***Poole et al.*** A study at the Universities of Auckland and Otago[66], in New Zealand (N=1,346), in which UMAT was also studied. Admission into medical school is based on performance during the first year at university, primarily but not entirely on grade point average. Overall 'admission GPA' accounted for a median of 11.7% of variance in medical school performance in years 2 to 6 (median r = .342).

2. Predictive validity of educational attainment: University in general.

a. ***Early studies in the United States.*** Studies in the United States have looked at correlations between high-school educational achievement and college attainment throughout most of the twentieth century. Fishman and Pasanella in 1960[67] reported that for 263 studies there was an average correlation with attainment in higher education of .50 for high-school grades or ranks, while for aptitude tests such as the SAT there was an average correlation of .47. The authors also noted, interestingly, that "Group intelligence tests … were less commonly employed, because they have proved less satisfactory than tests geared more directly to the measurement of scholastic abilities" (p.300). Somewhat earlier, Garrett in 1949[68] reviewed studies of high-school average as a predictor of performance at "Colleges of Arts and Science and Teachers Colleges"[kk],and reported a median correlation of .56. Garrett also cited the work of Douglass (1931), who reviewed 67 studies, with a mean correlation of 0.54, and Symonds (1929), who found an average correlation of 0.47 for 28 studies. Douglass concludes his study by suggesting that High School records had the highest correlation with College attainment (0.56), followed by General Achievement Test scores (.49), followed by intelligence tests (,47), and General College Aptitude tests (.43)[ll].

b. ***Berry and Sackett's analysis of US SAT data.*** A very large and very careful re-analysis was carried out by Berry and Sackett[69] of College Board data collected between 1995 and 1995 of 167,816 students entering 41 US colleges on a range of different courses.

[jj] Only a very weak measure of Finals Outcome was available, dividing students into pass and fail (with the vast majority passing); nevertheless A-levels correlated .063 with finals (p<.001, n=2510), but there was no correlation with AH5 total (p=.718, N=616).

[kk] These are very varied in their status, and include some of the best universities in the US.

[ll] The correlation with intelligence tests is rather higher than that reported more generally and may reflect a wider range of performance according to what counts as 'College'. Either way, given the stated high correlation of intelligence with high school performance, it is unlikely that it contributes much, if perhaps any, incremental variance.

For 9,889 courses the mean correlation of freshman grades with High School GPA was .293 (95% CI = .290 to .296)[mm].

c. ***The ISPIUA study.*** This study[58,59], to be considered in more detail below, looked at 5,985 entrants to university in the UK in the late 1960s and found a correlation of 0.32 between first year performance and mean A-level grade.

d. ***Bagg's 1970* Nature *study*.** This paper has been influential and much cited, mainly for its "principal conclusion that A-level grades of pass form an unreliable and possibly hazardous assessment, or prediction of the future academic performance, of a candidate", but in part also because it appeared in *Nature*. The major result is for 621 Honours students studying Chemical Engineering, where the median correlation between Chemistry, Physics and Maths A-levels, and four sets of marks (Mid-Sessional, Sessional, Part I finals and Part II finals) was .313 (range = .206 to .418), the median being remarkably close to the more general ISPIUA finding above. Other results were given for civil engineering, mechanical engineering, geography and history, although samples were much smaller, and extracting correlations from the regression equations they report is far from straightforward.

e. ***Naylor and Smith Study***. This study by Naylor and Smith[70], published in 2002, considered data from the Universities Statistical Record (USR; this later became the Higher Education Statistics Agency, HESA), on 48,281 undergraduates leaving university in 1992/3. A complex ordinal logistic regression was carried out on degree class, and there was a highly significant effect of A-level grades with dummy variables for males and females based on the top score being .458 and .350, both p<.001. There was also a highly significant effect whereby students schooled in the independent sector underperformed relative to the A-level grades they had attained. Extracting a simple correlation to describe the relationship of A-level grades to outcome is not straightforward. However, on p.6 we are told that 82.9% (85.7%) of male (female) students with 30 points (AAA) gain a good degree (I/II.i) compared with 54.4% (66.4%) of students gaining 22 to 25 points (about BBB).

f. ***HEFCE Study.*** In 2003, HEFCE published an analysis of 18-year olds entering UK universities in 1997/8, who had A-level grades, and graduated in 2001/2[14]. Detailed statistics and analyses are available online[nn]. For 79,005 students, it is possible to calculate the correlation between A level grades and university performance. The resulting correlation coefficient is 0.390. Restricting the analysis only to the 27,601 students with 24 points (BBB) or above, the correlation is reduced to 0.209, which is the expected deduction in correlation due to the effect of restriction of range.

g. ***The Cambridge Study***. This study[30] looked at correlations of various measures with examination results in Cambridge entrants taking Part 1 exams in the years 2006-9 (median N=540, range=304 to 1655). A-levels were not reported. The Pearson correlation of exam outcome in twelve independent groups across ten different subjects with best three AS grades had a median of .38 (range=.22 to .48), and the correlation with GCSE grades (number of A* results) had a median of .27 (range=.13 to .32).

h. ***The Sutton Trust Study***. This study[71], discussed further below, found a correlation of 0.38 between average A-level points score and degree class outcome (n= 2,754). The correlation with GCSE points was.36.

[mm] It is noteworthy that the correlation with high school GPAs is higher than for SATs, particularly given that the correlation of SATs with High GPAs was .253, suggesting that SATs have little additional variance once high school achievement has been taken into account.

[nn] http://www.hefce.ac.uk/pubs/hefce/2003/03_32.htm#desc

i. ***Chapman's study of USR data.*** Chapman[72] studied the relationship between A-levels and degree class in 254,402 students. Unfortunately though, all of the results presented are aggregated at the level of subject and university, and no simple raw correlations are provided. Having said that, across eight subjects, the correlation at the level of the university varied from 0.475 in Biology to 0.230 in Politics, with a mean of 0.391 (SD .098), which is very similar to correlations reported at the individual level of analysis.

j. ***The meta-analysis of Richardson et al.*** Richardson et al [52] carried out a large meta-analysis of factors predicting performance in higher education, mainly emphasising a range of "non-intellective" (i.e. personality and study habit) variables. However their systematic review of the literature for studies between 1997 and 2010 found only four studies which looked at A-levels in relation to university outcome, based on a total of 933 participants, which found a simple correlation of .25 with university outcome (disattenuated correlation = .31). They also found 46 studies looking at high school GPA (n=34,724), with a simple correlation with outcome of .40 (disattenuated = .41).

3. Predictive validity of aptitude tests: Medical school studies

a. ***UKCAT.*** We are aware of only three studies that have examined the predictive validity of UKCAT, all of which are fairly small.

i) **Lynch et al.** Lynch et al[73] looked at all 292 Year 1 entrants to Dundee or Aberdeen in 2007 with UKCAT and performance data. Overall the Pearson correlation was .062, which was not significant, and no account was taken of Scottish Higher or A-level results.

ii) **Wright and Bradley**. Wright and Bradley[74] looked at 307 students entering medical school in 2007 and 2008 on a range of examinations. Although they state that, "UKCAT was a significant predictor on all but one knowledge examination", no estimate of the correlation seems possible from the data presented[oo], and A-levels were not taken into account.

iii) **Yates and James**. Yates and James[75] studied 204 entrants to Nottingham Medical School in 2007 who had taken UKCAT. Results in four different 'Themes' were averaged over the first two years, and correlated .211, .126, .232, -.085 and -.014 (mean = .094) with UKCAT total score; the authors concluded, "the predictive value of the UKCAT … is low in the pre-clinical course at Nottingham".

b. ***BMAT.*** BMAT (BioMedical Admissions Test), which is currently required by four UK medical schools (Oxford, Cambridge, Imperial College and UCL), as well as by two Veterinary Schools, has three sections, Section 1 (Aptitude and Skills), Section 2 (Scientific Knowledge and Interpretation), and Section 3 (Writing Task). Section 3 is only used qualitatively[76]. Section 2 can be regarded primarily as a test of attainment, whereas Section 1 is a test of cognitive ability. There are two studies of BMAT, both in Cambridge with separate cohorts.

i) **Emery and Bell (2009).** The first published evidence on the predictive validity of BMAT[76], studied 1002 students who took the test (or its predecessor MVAT) between

[oo] Although their Table 2 says results are beta values, that is clearly not the case, with a constant of, say, 14.67 for Year 1 November. The UKCAT results are presented for that same group as "0.02 (0.00)", the figure in parentheses being the standard error. Despite much struggling it does not seem possible to get any straightforward estimate of effect size from these data.

2000 and 2003, and were followed into the second year of the course[pp]. Meta-analytic summaries of the various correlations[77] found correlations for first and second year courses with Section 1 of .19 and .15 and with Section 2 of .36 and .23. Partial correlations suggested that only Section 2 had predictive validity which was independent of the other section[77]. BMAT has particularly been criticised for a lack of evidence on incremental validity[77-79]. BMAT has defended itself on the grounds that almost all Cambridge students have three A grades at A-level and hence there is little or no variance in the educational attainment of medical school applicants. That explanation is not, though, entirely compatible with the more recent report, which follows.

ii) **The 2011 Cambridge study**. This study from Cambridge[30], which looked at a wide range of subjects and included 1881 students studying medicine, reported correlations of Part 1 and Part 2 outcome with AS-levels and GCSEs of .38 and .27, implying that not all students are at ceiling. The University of Cambridge website describing the study had the simple title, "A levels are a strong indicator of degree potential"[qq]. The mean simple correlations of outcome with BMAT 1 and BMAT 2 were .19 and .26.

c. ***UMAT.*** UMAT (Undergraduate Medical and Health Sciences Admission Test) consists of three sections, 1 (Logical reasoning and problem solving), 2 (Understanding people), and 3 (Non-verbal reasoning). The first and last sections are therefore concerned with general cognitive ability. The second section deals with what is believed to be a specific ability of particular relevance to the job role in question. UMAT is used mainly in Australasia for selecting medical students who are school-leavers.

(1) **Wilkinson et al.** In a study[80] of the predictive validity of UMAT at the University of Queensland (N=339), medical school grades "showed only a weak correlation [of.15 with] UMAT overall score … Further, the weak correlation did not persist beyond the first year of university study, and in multivariate analysis, correlation was limited to UMAT Section 1 score".

(2) **Merver and Puddey**. In a careful study at the University of Western Australia[61] (N=1174), which looked at cohorts of students entering from 1999 to 2009, educational attainment (TER) correlated .468, .401, .321, .230, .208 and .206 with grades in years 1 to 6 respectively, whereas UMAT total score correlated .030, -.001, -.006, .035, .038 and -.044 with the year grades (and none of the subscores performed any better). Overall it would seem that UMAT has little predictive validity.

(3) **Poole et al**. A study at the Universities of Auckland and Otago[66], in New Zealand (N=1346), which looked at outcomes in year 2 through to year 6. The statistics were varied and complex, but the overall UMAT score accounted for a median of 2.45% of the variance (r=.156). The study also concluded that, “As the predictive power of the UMAT score is so low, we cannot draw firm conclusions about the relative predictive power of individual UMAT sections”.

d. ***GAMSAT.*** The Graduate Medical School Admissions Test (GAMSAT) has three parts, I (Reasoning in Humanities and Social Sciences), 2 (Written Communication), and 3 (Reasoning in Biological and Physical Sciences), section 3 being an achievement test, while

[pp] Unusually and importantly there is a small scale follow-up of the 2000 and 2001 entrants into the clinical course, although only for those studying in Cambridge and only for the Pathology exams taken in the first clinical year (N=100 and N=113); with the written exam, sections 1 and 2 correlated .238 and .285 for the 2000 cohort and .147 and .311 for the 2001 cohort, with no correlations at all with the OSCE exam in pathology.

[qq] http://www.cam.ac.uk/admissions/undergraduate/research/a_levels.html

sections 1 and 2 are mainly aptitude/ability tests. GAMSAT was developed in Australia and has mainly been used there, although several graduate schools in the UK and Ireland also use it.

(1) **Wilkinson et al**. At the University of Queensland[62], a study (N=706) found good correlations of year 1 and year 4 outcomes with grade-point average (GPA) (r_s = .50 and .35), and weaker correlations with GAMSAT total score (r_s =.25 and .06), with partial correlations for GAMSAT being only .11 and -.01 after taking GPA into account (while partial correlations for GPA remained at .45 and .36).

(2) **Groves et al.** In a small study[81] (N=189), Groves et al. found that GAMSAT total score had only weak correlations with Year 2 results (r=.23 and .18), with much of the correlation being with scores on the the Biological and Physical Sciences subscale. This sub-scale is clearly a test of attainment.

(3) **Quinlivan et al**. In a study[82] at the University of Notre Dame in Australia (N=223), while GPA showed significant correlations with all of the outcome measures (correlation with Final Marks = .19, p=.005), GAMSAT total mark did not show significant correlations with the measures (correlation with Final Marks = .10, p=.16), and only scores on the Biological Sciences subscale showed a significant correlation with outcome.

(4) **Coates.** In a complex study, Coates[63] looked at educational attainment and medical school performance in six unnamed Australian institutions. Only 65% of the students who were invited to take part did so, yielding a total sample size of 351. Year 1 grades correlated significantly with previous GPAs (r=.47), and with total GAMSAT (median from Table 2 = .28), most of the effect being with GAMSAT 3 (median = .27), but not with GAMSAT 1and 2 (medians = .05 and .05). Compared with GPA alone, GAMSAT increased the variance accounted for from 22% to 28%.

(5) **Blackman and Darmawan.** A very small study[83] (n=99), also in Australia, found no statistically significant relationship of GAMSAT to outcome.

(6) **Bodger et al.** A recent small study[65] (N=105) looked at GAMSAT in graduate entrants to the University of Wales. There was a significant relationship of GAMSAT to overall pre-clinical performance (p<.001), although it is not possible to obtain a simple correlation for the strength of the relationship from the published data. However, the authors of the study have kindly sent us the unpublished correlations with first year exam results, which are .312, -.002 and .418 for GAMSAT 1, 2 and 3 (N=102, p=.001, NS, p<.001). We know of no other published studies on the use GAMSAT in the UK and Ireland.

e. ***ISPIUA and TAA***. The IPSIUA project[59] included 294 students studying medicine. At the end of the first year the Maths and Verbal tests of the TAA (Test of Academic Aptitude) correlated .10 and .15 (median =.12) with first year assessment.

f. ***Intelligence tests*** The AH5 and Raven's Progressive Matrices are conventional timed tests of intelligence (which can be regarded as tests of general cognitive ability having been shown in many studies to be predictive of both job-performance and 'trainability'[84]).

i) **AH5.** The AH5 test is a timed test of intelligence, developed by Alice Heim in the 1960s as a 'high-grade' test particularly suitable for adults at the high-end of the range, as in university students[85]. It can be regarded as a pure test of cognitive ability. Two studies that we know of have examined the extent to which it predicts performance in medical students.

(1) **The Westminster follow-up.** One of us looked at 511 clinical students who entered the Westminster Medical School between 1975 and 1982[13], and A-levels predicted performance in the clinical years and, of particular interest, the rate of gaining Membership exams and achieving Consultant/Principal status, whereas the AH5 showed no independent predictive ability.

(2) **The 1991 cohort study.**An abbreviated AH5 was also used in the 1991 cohort study carried out by one of us[64], and those data were also analysed for the 2005 paper on aptitude tests[13], and are provided in its online supplementary information. A-level results correlated .202 with Basic Medical Sciences results (n=3112, p<.001), and with mark at Part 1 of MRCP (n=903, p<.001), whereas AH5 total mark correlated only .044 (NS, N=766) and .123 (p=.057, N=240) with Basic Medical Science and MRCP Part 1 results[rr].

ii) **Raven's Progressive Matrices.** In an unusual study, Raven's Progressive Matrices, a non-verbal test of general cognitive ability, was used in a very small sample of 68 medical students at Beersheva in Israel[86]. Neither the correlation with medical school outcomes (median r = -.11), nor the correlation with school matriculation scores (median r = .10), was statistically significant. Perhaps the most striking feature of this study was its focus on a test specifically designed to measure general cognitive ability rather than one designed to measure 'aptitude'; the latter composed of one or more sub-scales concerned with general cognitive ability (e.g. verbal ability and abstact reasoning ability) and one or more sub-scales concerned with attainment (e.g. scientific knowledge).

g. ***MCAT.*** MCAT (Medical College Admission Test) is the standard admission test for American medical schools. It currently has four sub-scales (Physical Sciences, Biological Sciences, Verbal Reasoning and Writing Sample), of which the first two and possible the last, are attainment tests, and the third is a measure of general cognitive ability.

i) **MCAT meta-analysis**. In a meta-analysis[87] in 2007, the Biological Sciences subtest had good predictive validity, both for medical school measures (r=.40) and Licensing Examinations (Step 1 and Step 2, r=.58 and .38), while the Physical Sciences test and the Verbal Reasoning subtest have lower validity (r=.26 and .24 for medical school measures, and .52/.28 and .34/.34 for Step 1/2). The Writing Sample has little predictive validity. No validity coefficients were provided for subtests, taking into account performance on other subtests, but it is likely that Biological Sciences is responsible for most of the effect.

ii) **Callahan et al.** A recent study by Callahan et al[88] looked at the performance of 7,859 students who matriculated at Jefferson Medical College from 1970 to 2005, and who had taken three different versions of MCAT. The average correlation of the total score with first and second year results was 0.32. Correlations with first year results were not presented by sub-scores, but correlations with NBME Part 1 or USMLE Part 1 were .41 for science sub-scores, .28 for quantitative sub-scores (pre-1992), and .27 for verbal reasoning/reading. As in the meta-analysis, Biological Sciences correlated more highly than Physical Sciences (.44 vs .36).

h. ***TMS and EMS***. TMS (Tests für medizinische Studiengänge) and EMS (Eignungstest für das Medizinstudium in der Schweiz) are used for selection into medical schools in Germany

[rr] Only a very weak measure of Finals Outcome was available, dividing students into pass and fail (with the vast majority passing); nevertheless A-levels correlated .063 with finals (p<.001, n=2510), but there was no correlation with AH5 total (p=.718, N=616).

(TMS) and in Switzerland and Austria (EMS). They have a range of different forms of test material, much of which is based in scientific analysis and education, as well as measures of memory and spatial ability, and to a large extent appear to be more measures of attainment than general cognitive ability[ss]. A recent meta-analysis[89] of these and other German-language subject-specific tests, which looked at 36 independent samples involving 45,091 participants, across all subject-areas found an uncorrected correlation of .321 with first year course outcome[tt]. The correlation for the 34,438 participants studying medicine was higher at about .340 (95% CI .323 - .357)[uu].

4. Predictive validity of aptitude tests: University in general

a. ***ISPIUA and TAA***. The use of aptitude tests for selection to university in general in the UK goes back to the ISPIUA project (Investigation into Supplementary Predictive Information for University Admission), which in 1967 administered the Test of Academic Aptitude (TAA) to 27,000 sixth-formers, following them up a few years later at the end of their first year of university [59]. For 5,985 students, the correlations of the Maths and Verbal subtests with first year assessments were .01 and .14 [p.29], and the multiple R was .12 [p.32][vv]. The same students were followed up to the end of their (three-year) degree courses (n=4175), and degree class correlated -.02 and .13 with M and V, and multiple R = .13 [pp.47, 49]. A further analysis, in not so much detail, was carried out of the 1,408 students who entered university in 1969[ww]. M and V correlated .11 and .17 with first year assessments (multiple R = .18). Taking the results overall, there was evidence of a correlation of TAA with university outcomes, although most of the variance was already accounted for by A-levels and O-levels. The results were summarised succinctly by James Drever, a one-time advocate of the test when he sat on the Robbins Committee, "test scores improve prediction only marginally when used in conjunction with A- and O-level grades" [90,91].

b. ***The Sutton Trust SAT study***. Forty years after ISPIUA, The Sutton Trust, which has advocated the use of an American SAT test for university admission, commissioned an extensive and useful literature review [92], which also referred to the only major British study, ISPIUA and the TAA, and concluded TAA, "added virtually nothing to the prediction of degree grades in addition to A-levels" (p.6). The Sutton Trust nevertheless commissioned a major empirical study, administering the American SAT in autumn 2005 to over 9000 sixth-formers, 2,754 of whom were followed up when they graduated in 2009. The correlation of degree outcome with mean SAT score was 0.26, correlations with the subscores being .26, .24 and .18 for Writing, Reading and Maths. The conclusion was similar to that of ISPIUA, in that "average A-level points is the best predictor of HE participation and degree class", and that, "In the absence of other data, the SAT has some predictive power but it does not add any additional information, over and above that of GCSEs and A-levels (or GCSEs alone), at a significantly useful level" [p.1][71].

c. ***The Cambridge analysis***. In 2011 the University of Cambridge put on its website a brief report describing the correlation of several aptitude tests, as well as AS-levels and GCSE grades, with first and second year examination results. Across 12 subject/year

[ss] For a complete example paper from TMS see http://www.tms-info.org/content/files/informationsbroschuere_tms2012.pdf , and for other information, including a list of participating medical schools see http://www.tms-info.org/ and http://www.unifr.ch/ztd/ems/ .

[tt] When corrected for unreliability of the tests and the outcome measures, the disattenuated correlation was .478.

[uu] The uncorrected correlation is not given, but the disattenuated correlation is .507 with a 95% CI of .481 to .532), and the value quoted in the text has been back calculated from those figures.

[vv] The report comments that, "for virtually all courses, this multiple correlation is only slightly different from the larger of the correlations with each of these variables separately" [p.33].

[ww] ICM was one of the participants in this study.

combinations (10 subjects), AS-levels on average correlated .38 with outcome (SD=.07, range .22 to .48), and GCSEs correlates on average 0.26 with outcome (SD=.06, range=.13 to .32). Amongst the aptitude tests, STEPII and STEP III, which are to a large extent mathematical attainment tests, correlated .47 and .54 with outcome, BMAT 1 (which is more of a general cognitive ability test) correlated .17, and BMAT 2 (which is more of an attainment test) correlated .24 with outcome, and finally TSA (Thinking Skills Assessment with Critical Thinking and Problem Solving subtests), correlated.15 and .17 in the four subjects (SD .05 and .08; range = .08 to .19 and .10 to .25). It seems clear that as tests become less of an attainment test, with specific subject content, and more of a test of general cognitive ability, so the correlation with outcome falls. The analysis concludes, albeit with some provisos, that "Aptitude tests have been a less effective predictor of Tripos performance overall … but the BMAT has had a positive utility". The claim of ‘positive utility’might be regarded as contentious and probably only applies, if at all, to BMAT 2[77].

d. ***The Poropat (2009) meta-analysis*.** Measures of ability/aptitude are often very similar in content and structure to tests of intelligence (IQ, general mental ability), and it therefore makes sense to compare specific aptitude/ability tests with conventional tests of general cognitive ability, which are typically very reliable. Although the primary interest of the meta-analysis of Poropat[93] was in the relationship between academic performance and personality measures, the analysis included 26 studies in which academic performance at tertiary level and general cognitive ability had been studied in 17,588 individuals. A disattenuated correlation of .23 was reported[xx], and given typical reliabilities of general cognitive ability tests of of .9 and of grade-point average of about .84, the raw (attenuated) correlation is about 0.20. The paper does not give a 95% confidence interval, but given the sample size, it is likely to be small.

e. ***Berry and Sackett’s analysis of US SAT data*.** As mentioned earlier, a very large and very careful re-analysis was carried out by Berry and Sackett[69] of College Board data collected between 1995 and 1995 from 167,816 students entering 41 US colleges on a range of different courses. An overall SAT score (combined verbal and maths subtests) was examined in relation to performance on individual courses[yy]. For 10,117 courses the mean correlation of freshman grades with SATs was .257 (95% CI.253 to .261).

f. ***Hell et al, 2007*.** Hell et al. [89] reported a meta-analysis of German-language, subject-specific aptitude tests used in Germany, Switzerland and Austria. Although a range of disciplines was included, 25 of the 36 studies were for medicine, veterinary medicine or dentistry, putting in doubt the generalizability of the overall result.

g. ***The meta-analysis of Richardson et al.*** The meta-analysis of Richardson et al[52] mentioned earlier, described 35 studies with 7820 participants which had looked at measures of intelligence in relation to outcome of higher education. The raw correlation was 0.20, with a disattenuated correlation of 0.21. That correlation is very similar to the value of .23 reported by Poropat et al in tertiary education.

[xx] It is noteworthy that there is a strong trend, as many researchers have suspected, for intelligence tests to correlate less well with academic achievement at higher educational levels, and Poporat finds that also, the correlation being .57 at primary level, compared with .23 at tertiary level.

[yy] Although most studies look at overall GPA, there are a number of problems with that, and individual course grades within colleges have many advantages, and also are more directly comparable with the UKCAT data, where results are within medical schools.

Reference List

1. UKCAT. *UKCAT: 2006 Annual report*. Cambridge: United Kingdom Clinical Aptitude Test (available at http://www.ukcat.ac.uk/pdf/UKCAT Annual Report 2006 v2.pdf); 2008.

2. Hunter JE. Cognitive-ability, cognitive aptitudes, job knowledge, and job-performance. 1986;29:340-62.

3. Olea MM, Ree MJ. Predicting pilot and navigator criteria - not much more than g. 1994;79:845-51.

4. Ree MJ, Earles JA, Teachout MS. Predicting Job-Performance - Not Much More Than G. *J.Appl.Psychol.* 1994;79:518-24.

5. Schmidt FL, Ones DS, Hunter JE. Personnel-Selection. *Annu.Rev.Psychol.* 1992;43:627-70.

6. Brand C. *The g factor: General intelligence and its implications*. New York: Wiley; 1996.

7. Gottfredson LS. Why g matters: The complexity of everyday life. *Intelligence* 1997;24:79-132.

8. Jensen AR. *The G factor - The science of mental ability*. Westport, CT: Praeger; 1998.

9. Reeve CL, Hakel MD. Asking the right questions about g. 2002;15:47-74.

10. Ree MJ, Earles JA. Predicting training success - Not much more than g. *Pers Psychol* 1991;44:321-32.

11. Ree MJ, Carretta TR. g2K. 2002;15:3-23.

12. Schmidt FL. The role of general cognitive ability and job performance: Why there cannot be a debate. *Hum Perform* 2002;15:187-210.

13. McManus IC, Smithers E, Partridge P, Keeling A, Fleming PR. A levels and intelligence as predictors of medical careers in UK doctors: 20 year prospective study. *British Medical Journal* 2003;327:139-42.

14. Anonymous. *Schooling effects on higher education achievement*. http://www.hefce.ac.uk/pubs/hefce/2003/03_32.htm: 2003.

15. Cassidy J. UKCAT among the pigeons. *British Medical Journal* 2011;336:691-2.

16. McManus IC, Powis DA, Wakeford R, Ferguson E, James D, Richards P. Intellectual aptitude tests and A levels for selecting UK school leaver entrants for medical school. *British Medical Journal* 2005;331:555-9.

17. NcLennan D, Barnes H, Noble M, Davies J, Garratt E, Dibben C. *The English Indices of Deprivation 2010*. London: Department for Communities and Local Government; 2011.

18. Emery JL, Bell JF, Emer, Rodeiro CLV. The BioMedical Admissions Test for medical student selection: Issues of fairness and bias. *Medical Teacher* 2011;33:62-71.

19. Bacon DR, Bean B. GPA in research studies: An invaluable but neglected opportunity. *Journal of Marketing Education* 2006;28:35-42.

20. Tighe J, McManus IC, Dewhurst NG, Chis L, Mucklow J. The Standard Error of Measurement is a more appropriate measure of quality in postgraduate medical assessments than is reliability: An analysis of MRCP(UK) written examinations. *BMC Medical Education (www.biomedccentral.com/1472-6920/10/40)* 2010;10:40.

21. Rosenthal R. How are we doing in soft psychology? *American Psychologist* 1990;45:775-7.

22. UKCAT. *UKCAT: 2007 Annual Report*. Cambridge: United Kngdom Clinical Aptitude Test (available at http://www.ukcat.ac.uk/pdf/UKCAT Annual Report 2007.pdf); 2008.

23. UKCAT. *UKCAT 2008 Annual Report*. Nottingham: UKCAT; 2008.

24. Pearson Vue. *Technical Report: Testing Interval: July 11 2006 - October 14,2006 (January 2007, Revised May 3, 2007)*. Chicago: Pearson Vue; 2007.

25. Pearson Vue. *Technical Report regarding the UK CLinical Aptitude Test Examination for the UKCAT consortium; Testing Interval: 18 June 2007 - 10 October 2007; January 2008*. Chicago: Pearson Vue; 2008.

26. Pearson Vue. *Technical Reports; UK Clinical Aptitude Test Consortium; UKCAT examination; Teseting Interval: 7 July 2008 - 12 October 2008*. Chicago: Pearson Vue; 2009.

27. Johnson J, Hayward G. *Expert group report for award seeking admission to the UCAS tariff: Scottish Highers and Advanced HIghers*. Cheltenham: UCAS; 2008.

28. Richards P, Stockill S, Foster R, Ingall E. *Learning Medicine: How to become and remain a good doctor (18th edition)*. Cambridge: Cambridge University Press; 2011.

29. Woolf K, Potts HWW, McManus IC. The relationship between ethnicity and academic performance in UK-trained doctors and medical students: a systematic review and meta-analysis. *British Medical Journal* 2011;342:d901-doi: 10.1136/bmj.d901.

30. Partington R. *The Predictive Effectiveness of Metrics in Admission to the University of Cambridge*. Cambridge: University of Cambridge: http://www.cam.ac.uk/admissions/undergraduate/research/docs/prefective_effectiveness_of_metrics_in_admission.pdf; 2011.

31. McManus IC, Richards P. Prospective survey of performance of medical students during preclinical years. *Brit.Med.J.* 1986;293:124-7.

32. Arneson JJ, Sackett PR, Beatty AS. Ability-performance relationships in education and employment settings: Critical tests of the More-Is-Better and Good-Enough hypotheses. *Psychological Science* 2011;22:1336-42.

33. Woolf K, McManus IC, Gill D, Dacre JE. The effect of a brief social intervention on the examination results of UK medical students: a cluster randomised controlled trial. *BMC Medical Education* 2009;9:35.

34. Bekhradnia B, Thompson J. *Who does best at University?* London: Higher Education Funding Council England. **www.hefce.ac.uk/Learning/whodoes;** 2002.

35. James D, Yates J, Nicholson S. Comparison of A level and UKCAT performance in students applying to UK medical and dental schools in 2006: cohort study. *British Medical Journal* 2010;340:c478.

36. Garlick PB, Brown G. Widening participation in medicine. *British Medical Journal* 2008;336:1111-3.

37. Mahesan N, Crichton S, Sewell H, Howell S. The effect of an intercalated BSc on subsequent academic performance. *BMC Medical Education* 2011;11:76:www.biomedcentral.com/1472-6920/11/76.

38. McManus IC. To BSc or not to BSc ... *Student BMJ* 2011;19:d7559-10.1136/sbmj.d7559.

39. Cortina JM. What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology* 1993;78(1):98-104.

40. Schmitt N. Uses and abuses of coefficient alpha. *Psychological Assessment* 1996;8(4):350-3.

41. Thorndike RL. *Personnel selection*. New York: Wiley; 1949.

42. Hunter JE, Schmidt FL, Le H. Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology* 2006;91(3):594-612.

43. Le H, Schmidt FL. Correcting for indirect range restriction in meta-analysis: Testing a new meta-analytic procedure. *Psychological Methods* 2006;11(2006):416-38.

44. Ghiselli EE, Campbell JP, Zedeck S. *Measurement theory for the behavioral sciences*. San Francisco: W H Freeman; 1981.

45. Stauffer JM, Mendoza JL. The proper sequence for correcting correlation coefficients for range restriction and unreliability. *Psychometrika* 2001;66(1):63-8.

46. McManus IC, Woolf K, Dacre J. The educational background and qualifications of UK medical students from ethnic minorities. *BMC Medical Education* 2008;8: 21 (http://www.biomedcentral.com/1472-6920/8/21).

47. He Q. *Estimating the reliability of composite scores*. Coventry: Office of Qualifications and Examinations Regulation (http://www.ofqual.gov.uk/files/2010-02-01-composite-reliability.pdf); 2009.

48. Opposs D, He Q. *The reliability programme: Final Report*. Coventry: Office of Qualifications and Examinations Regulation (http://www.ofqual.gov.uk/files/reliability/11-03-16-Ofqual-The-Final-Report.pdf); 2011.

49. Wheadon C, Stockford I. *Classification accuracy and consistency in GCSE and A level examinations offered by rthe Assessment and Qualifications Alliance (AQA) November 2008 to June 2009*. Office of Qualifications and Examinations Regulation: Coventry (http://www.ofqual.gov.uk/files/reliability/11-03-16-AQA-Classification-Accuracy-and-Consistency-in-GCSE-and-A-levels.pdf); 2011.

50. Bramley T, Dhawan V. *Estimates of reliability of qualifications*. Coventry: Office of Qualifications and Examinations Regulation (http://www.ofqual.gov.uk/files/reliability/11-03-16-Estimates-of-Reliability-of-qualifications.pdf); 2011.

51. Sheng Y, Sheng Z. Is coefficient alpha robust to non-normal data? *Frontiers in Psychology* 2012;34(February):#34.

52. Richardson M, Abraham C, Bond R. Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin* 2012;138(2):353-87.

53. Yates J. *Report on analysis of UKCAT scores and school qualifications data (2006 cohort).* Unpublished report: 2006.

54. McManus IC, Woolf K, Dacre JE. Even one star at A level could be "too little, too late" for medical student selection. *BMC Medical Education* 2008;8:16 (http://www.biomedcentral.com/1472-6920/8/16).

55. Woolf K. *The academic underperformance of medical students from ethnic minorities (unpublished PhD thesis).* London: University College London; 2012.

56. Ferguson E, James D, Madeley L. Factors associated with success in medical school and in a medical career: systematic review of the literature. *British Medical Journal* 2002;324:952-7.

57. Cohen-Schotanus J, Muijtjens AMM, Reinders JJ, Agsteribbe J, Van Rossum HJM, van der Vleuten CPM. The predictive validity of grade point average scores in a partial lottery medical school admission system. *Medical Education* 2006;40:1012-9.

58. Choppin BHL, Orr L, Kurle SDM, Fara P, James G. *The prediction of academic success.* Slough: NFER Publishing Company Ltd; 1973.

59. Choppin B, Orr L. *Aptitude testing at eighteen-plus*. Windsor: NFER; 1976.

60. Arulampalam W, Naylor RA, Smith JP. A Hazard Model of the probability of medical school dropout in the United Kingdom. *Journal of the Royal Statistical Society A* 2004;167:157-78.

61. Merver A, Puddey IB. Admission selection criteria as predictors of outcomes in an undergraduate medical course: A prospective study. *Medical Teacher* 2011;Early Online:1-8.

62. Wilkinson D, Zhang J, Byrne GJ, Luke H, Ozolins IZ, Parker MH et al. Medical school selection criteria and the prediction of academic performance: Evidence leading to change in policy and practice at the University of Queensland. *Medical Journal of Australia* 2008;188:349-54.

63. Coates H. Establishing the criterion validity of the Graduate Medical School Admissions Test (GAMSAT). *Medical Education* 2008;42:999-1006.

64. McManus IC, Richards P, Winder BC. Intercalated degrees, learning styles, and career preferences: prospective longitudinal study of UK medical students. *British Medical Journal* 1999;319:542-6.

65. Bodger O, Byrne A, Evans PA, Rees S, Jones G, Cowell C et al. Graduate entry medicine: Selection criteria and student performance. *PLoS ONE* 2011;6:11-doi:10.1371/journal.pone.0027161.

66. Poole P, Shulruf B, Rudland J, Wilkinson T. Comparisoon of UMAT scores and GPA in preduction of performance in medical school: A national study. *Medical Education* 2012;46:163-71.

67. Fishman JA, Pasanella AK. College admission-selection studies. *Review of Educational Research* 1960;30:298-310.

68. Garrett HF. A review and interpretation of investigations of factors related to scholastic success in colleges of arts and sciences and teachers colleges. *Journal of Experimental Education* 1949;18:91-138.

69. Berry CM, Sackett PH. Individual differences in course choice result in underestimation of the validty of college admissions sytems. *Psychological Science* 2009;20:822-30.

70. Naylor R, Smith J. *Schooling effects on subsequent university performance: Evidence for the UK university population*. Warwick: Warwick Economic Research Papers No 657: www2.warwick.ac.uk/fac/soc/economics/research/papers/twerp657.pdf; 2002.

71. Kirkup C, Schagen I, Wheater R, Morrison J, Whetton C. *Use of an aptitude test in university entrance - A validity study: Relationships between SAT scores, attainment measures and background variables*. London: Department for Education and Skills, Research Report RR846; 2007.

72. Chapman K. Entry qualifications, degree results and value-added in UK universities. *Oxford Review of Education* 2011;22:251-64.

73. Lynch B, MacKenzie R, Dowell J, Cleland J, Prescott G. Does the UKCAT predict Year 1 performance in medical school? *Medical Education* 2009;43:1203-9.

74. Wright SR, Bradley PM. Has the UK Clinical Aptitude Test improved medical student selection? *Medical Education* 2010;44:1069-76.

75. Yates J, James D. The value of the UK Clinical Aptitude Test in predicting pre-clinical performance: A prospective cohort study at Nottingham Medical School. *BMC Medical Education* 2010;10:55-www.biomedcentral.com/1472-6920/10/55.

76. Emery JL, Bell JF. The predictive validity of the BioMedical Admissions Test for pre-clinical examination performance. *Medical Education* 2009;43:557-64.

77. McManus IC, Ferguson E, Wakeford R, Powis D, James D. Predictive validity of the BioMedical Admissions Test (BMAT): An evaluation and case study. *Medical Teacher* 2011;33:53-7.

78. Emery J, Bell JF. Comment on I.C. McManus, Eamonn Feguson, Richard WSakeford, David Powis and David James (2011). Predictive validity of the BioMedical Admissions Test (BMAT): An evaluation and case study. *Medical Teacher* 33(1) (this issue). *Medical Teacher* 2011;33:58-9.

79. McManus IC, Ferguson E, Wakeford R, Powis D, James D. Response to Comments by Emery and Bell, *Medical Teacher* 33(1): (this issue). *Medical Teacher* 2011;33:60-1.

80. Wilkinson D, Zhang J, Parker M. Predictive validity of the Undergraduate Medicine and Health Sciences Admission Test for medical students' academic performance. *Medical Journal of Australia* 2011;194:341-4.

81. Groves MA, Gordon J, Ryan G. Entry tests for graduate medical programs: Is it time to rethink? *Medical Journal of Australia* 2007;186:120-3.

82. Quinlivan JA, Lam LT, Wan SH, Petersen WR. Selecting medical students for academic and attitudinal outcomes in a Catholic medical school. *Medical Journal of Australia* 2010;193:347-50.

83. Blackman I, Darmawan IGN. Graduate-entry medical student variables that predict academic and clinical achievement. *International Education Journal* 2004;4:30-41.

84. Schmidt FL, Hunter JE. The validity and utility of selection methods in personnel psychology: practical and theoretical implications of 85 years of research findings. *Psychological Bulletin* 1998;124:262-74.

85. Heim AW. *AH5 group test of high-grade intelligence*. Windsor: NFER; 1968.

86. Hobfoll SE, Benor DE. Prediction of student clinical performance. *Medical Education* 1981;15:231-6.

87. Donnon T, Paolucci EO, Violato C. The predictive validity of the MCAT for medical school performance medical board licensing examinations: A meta-analysis of the published research. *Academic Medicine* 2007;82:100-6.

88. Callahan CA, Hojat M, Veloksi J, Erdmann JB, Gonnella JS. The predictive validity of three versions of the MCAT in relation to performance in Medical School, Residency, and Licensing Examinations: A longitudinal study of 36 classes of Jefferson Medical College. *Academic Medicine* 2010;85:980-7.

89. Hell B, Trapmann S, Schuler H. Eine Metaanalyse der Validität von fachspezifischen Studierfähigkeitstests im deutschsprachigen Raum. *Empirische Pädagogik* 2007;21:251-70.

90. McManus IC. Why I ... think introducing SATS for university applicants in Britain is a waste of time. *Times Higher Education Supplement* 2005;October 7th:16.

91. Whetton C, McDonald A, Newton P. *Aptitude testing for university entrance*. Slough: National Foundation for Education Research **www.nfer.ac.uk/research/papers/PaperforRio090301.doc;** 2001.

92. McDonald AS, Newton PE, Whetton C, Benefield P. *Aptitude testing for university entrance: A literature review*. London: National Foundation for Educational Research; 2000.

93. Poropat AE. A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin* 2009;135:322-38.

Table 1. Simple Pearson correlations of A-level, AS-level and GCSE measures with *OverallMark.* Note that N refers to 3 for A-levels, 4 for AS-levels and 9 for GCSEs. Key: * P<.05; ** P<.01; *** P<.001. Correlations with p<.05 are also shown in bold.

	A-levels	AS-levels	GCSE
Total number of exams	.005 (2725)	.024 (1842)	**.078 *** **(721)**
Points from N best grades	**.177 ***** **(2725)**	**.180 ***** **(1842)**	**.179 ***** **(721)**
Total points from all exams	**.077 ***** **(2725)**	**.108 ***** **(1842)**	**.158 ***** **(721)**
Biology taken	.030 (2725)	.009 (1842)	.047 (721)
Chemistry taken	.010 (2725)	-.007 (1842)	.043 (721)
Maths taken	.007 (2725)	.015 (1842)	**.073 *** **(721)**
Physics taken	**.048 *** **(2725)**	.034 (1842)	.061 (721)
General Studies taken (A- and AS- only)	**.051 **** **(2725)**	**.058 *** **(1842)**	n/a
Non-Science taken (A- and AS- only)	-.011 (2725)	-.004 (1842)	n/a
Number of non core-science (GCSE)	n/a	n/a	-.007 (721)
Double science taken (GCSE only)	n/a	n/a	-.044 (721)
Biology grade (if taken)	**.169 ***** **(2610)**	**.121 ***** **(1809)**	**.198 ***** **(379)**
Chemistry grade (if taken)	**.141 ***** **(2701)**	**.160 ***** **(1825)**	**.146 ***** **(379)**
Maths grade (if taken)	**.129 ***** **(1725)**	**.106 ***** **(1346)**	.030 (676)
Physics grade (if taken)	**.172 ***** **(674)**	**.187 ***** **(626)**	**.162 ***** **(.455)**
General Studies grade (if taken)	**.234 ***** **(713)**	**.159 ***** **(695)**	n/a
Double Science grade (GCSE only)	n/a	n/a	**.152 **** **(338)**

Table 2. The correlation matrix of the eight variables entered into the principal component analysis of A-level results. The upper triangle of the correlation matrix shows the simple Pearson correlation between the variables with N in parentheses, while the lower triangle shows the estimated correlation matrix with EM imputation for missing values. Principal component analyses are calculated separately for the EM imputed missing data and for mean value imputation, with the first column in each case indicating the loadings on the first principal component, and the coefficients for creating the factor scores. The final three columsn show descriptive statistics for the variables, with N indicating the number of valid cases (and hence an indication of the proportion of missing values). Overall N=2725.

	Correlation matrix (*Upper*: raw with N; *Lower:* EM substituted)								EM imputation		Mean value imputation		Descriptives		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	First principal component loadings	Component Score coefficients*	First principal component loadings	Component Score coefficients*	Mean	SD	N
(1) Alevel_TotalbestN Total score for best three Alevels	1	.418 (1842)	.298 (721)	.165 (713)	.558 (2610)	.685 (2701)	.483 (1725)	.439 (674)	.857	.237	.911	.388	29.3	1.31	2725
(2) ASlevel_TotalbestN Total score for best four ASlevels	**.479**	1	.416 (516)	.243 (529)	.211 (1774)	.316 (1826)	.224 (1157)	.240 (406)	.704	.195	.536	.228	38.3	2.59	1842
(3) GCSE_TotalbestN Total score for best nine GCSEs	**.429**	**.552**	1	.461 (196)	.256 (691)	.186 (719(	.140 (466)	.330 (150)	.752	.208	.295	.126	50.0	3.48	721
(4) Alevel_highest_GeneralStudies	**.262**	**.428**	**.667**	1	.186 (683)	.069 (708)	.097 (.042)	.350 (153)	.622	.172	.197	.084	8.21	2.13	713
(5) Alevel_highest_Biology	**.564**	**.282**	**.430**	**.370**	1	.198 (2589)	.142 (1617)	.035 (591)	.615	.170	.598	.255	9.79	0.64	2610
(6) Alevel_highest_Chemistry	**.687**	**.353**	**.291**	**.174**	**.213**	1	.151 (1717)	.182 (667)	.589	.163	.718	.306	9.60	0.87	2701
(7) Alevel_highest_Maths	**.557**	**.317**	**.254**	**.105**	**.171**	**.196**	1	.096 (447)	.550	.152	.410	.175	9.80	0.67	1725
(8) Alevel_highest_Physic	**.561**	**.298**	**.343**	**.328**	**.239**	**.212**	**.409**	1	.632	.175	.252	.107	9.55	1.00	674

Table 3. Simple Pearson correlations of Highers, HighersPlus and Advanced Highers with *OverallMark.* Note that N refers to 5 for Highers and Highers Plus, and 1for Advanced Highers. Key: * P<.05; ** P<.01; *** P<.001. Correlations with p<.05 are also shown in bold.

	Highers	'HighersPlus'	Advanced Highers
Total number of exams	-.207 (715)	-.041 (682)	**.160 ***** **(639)**
Points from N best grades	**.121 ***** **(715)**	**.248 ***** **(682)**	**.388 ***** **(639)**
Total points from all exams	.025 (715)	**.080 *** **(682)**	**.300 ***** **(639)**
Biology taken	.062 (715)	.060 (682)	.063 (639)
Chemistry taken	-.004 (715)	-.018 (682)	**.103 **** **(639)**
Maths taken	.038 (715)	.019 (682)	.000 (639)
Physics taken	.011 (715)	.004 (682)	.066 (639)
Number non-science exams taken	-.017 (715)	-.020 (682)	.056 (639)
Biology grade (if taken)	**.118 **** **(705)**	**.237 ***** **(667)**	**.363 ***** **(407)**
Chemistry grade (if taken)	.028 (711)	**.193 ***** **(677)**	**.467 ***** **(521)**
Maths grade (if taken)	**.076 *** **(704)**	**.120 **** **(672)**	**.268 ***** **(216)**
Physics grade (if taken)	**.113 *** **(481)**	**.200 ***** **(458)**	**.378 ***** **(116)**
Number of Core Sciences taken	.034 (715)	**.093 *** **(715)**	**.134 ***** **(715)**
Mean grade at Core Sciences	**.128 ***** **(715)**	**.248 ***** **(682)**	**.451 ***** **(633)**
Total points at Core Sciences	.064 (715)	**.107 **** **(682)**	**.293 ***** **(633)**
Maximum grade at a Core Science	-.031 (715)	**.095 *** **(682)**	**.416 ***** **(633)**
Minimum grade at a Core Science	**.167 ***** **(715)**	**.265 ***** **(715)**	**.423 ***** **(633)**

Table 4. The correlation matrix of the ten variables entered into the principal component analysis of SQA Highers. The upper triangle of the correlation matrix shows the simple Pearson correlation between the variables with N in parentheses, while the lower triangle shows the estimated correlation matrix with EM imputation for missing values. Principal component analyses are calculated separately for the EM imputed missing data and for mean value imputation, with the first column in each case indicating the loadings on the first principal component, and the coefficients for creating the factor scores. The final three columns show descriptive statistics for the variables, with N indicating the number of valid cases (and hence an indication of the proportion of missing values). Overall N=715.

	Correlation matrix (*Upper*: raw with N; *Lower:* EM substituted)										EM imputation		Mean value imputation		Descriptives		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	First principal component loadings	Component Score coefficients*	First principal component loadings	Component Score coefficients*	Mean	SD	N
(1) SQAhigherPlus_TotalbestN	1	.573 (667)	.668 (677)	.657 (672)	.467 (458)	.349 (628)	.410 (402)	.422 (513)	.267 (213)	.436 (112)	.805	.222	.739	.142	47.47	2.26	682
(2) SQAhigherPlus_highest_Biology	**.562**	1	.377 (662)	.288 (657)	.190 (444)	.331 (616)	.394 (398)	.335 (506)	.242 (206)	.286 (109)	.595	.164	.543	.105	9.52	.68	667
(3)SQAhigherPlus_highest_Chemistry	**.668**	**.371**	1	.488 (668)	.323 (456)	.331 (624)	.299 (402)	.423 (510)	.331 (212)	.393 (112)	.696	.192	.649	.125	9.55	.65	677
(4) SQAhigherPlus_highest_Maths	**.654**	**.295**	**.488**	1	.395 (449)	.260 (619)	.239 (396)	.379 (505)	.276 (213)	.434 (112)	.654	.180	.605	.116	9.47	.70	672
(5) SQAhigherPlus_highest_Physics	**.546**	**.223**	**.386**	**.464**	1	.289 (423)	.335 (239)	.362 (347)	.153 (160)	.307 (112)	.465	.128	.573	.110	9.48	.81	458
(6) SQAadvHigher_TotalbestN	**.382**	**.350**	**.359**	**.273**	**.301**	1	.797 (407)	.750 (521)	.725 (216)	.870 (116)	.709	.195	.836	.161	8.47	1.46	639
(7) SQAadvHigher_highest_Biology	**.452**	**.395**	**.368**	**.286**	**.349**	**.814**	1	.604 (333)	.357 (114)	.479 (41)	.572	.158	.824	.159	8.05	1.52	407
(8) SQAadvHigher_highest_Chemisty	**.431**	**.355**	**.417**	**.390**	**.366**	**.697**	**.641**	1	.389 (163)	.696 (88)	.666	.184	.790	.152	7.85	1.54	521
(9) SQAadvHigher_highest_Maths	**.325**	**.136**	**.289**	**.235**	**.198**	**.674**	**.609**	**.438**	1	.498 (41)	.336	.093	.659	.127	8.06	1.78	216
(10) SQAadvHigher_highest_Physics	**.453**	**.369**	**.433**	**.430**	**.438**	**.892**	**.792**	**.769**	**.664**	1	.333	.092	.898	.173	8.52	1.43	116

Table 5. Simple Pearson correlations of key measures with a range of demographic, school, social and UKCAT process measures. **Note**: measures *in italics* are **contextual measures**, and should be treated with care as they describe the student's environment rather than the student themselves. **Key**: * P<.05; ** P<.01; *** P<.001. Correlations with p<.05 are also shown in bold.

		Educational Attainment *zEducational Attainment*	3 /5 best A-levels/ Highers	UKCAT total score *zUKCATtotal*	Overall medical school score *OverallScore*
Demographic measures	UK national	.008 (3432)	.000 / -.042 (2764/ 769)	.060 *** (4811)	-.007 (4811)
	Male	-.037 * (3432)	.026 / .058 (2764 / 769)	.061 *** (4742)	-.039 ** (4742)
	Aged 21+	n/a	n/a	-.060 *** (4766)	.080 *** (4766)
	Aged 30+	n/a	n/a	-.023 (4766)	-.003 (4766)
	Ethnic2 (non-White)	-.053 ** (3221)	-.062 ** / -.033 (2549 / 766)	-.141 *** (4149)	-.142 *** (4149)
School measures	Selective Schooling	.051 ** (3432)	.038 * / .120 *** (2764 / 769)	.075 *** (4811)	-.101 *** (4811)
	DfES Value Added KS 5	-.012 (2092)	.012 / n/a (2119)	-.014 (2561)	-.049 * (2561)
	DfES Average points per student	.085 *** (2114)	.127 *** / n/a (2141)	.097 *** (2586)	-.065 *** (2586)
	DfES Average points per exam entry	.111 *** (2109)	.101 *** / n/a (2136)	.044 * (2582)	-.111 *** (2582)
Social background	Socio-economic classification (SEC) (1=High 5=Low)	-.058 * (2939)	-.084 *** / -.046 (2356 / 675)	-.056 *** (4091)	-.011 (4091)
	Overall Deprivation Decile *(1= high, 10 = low deprivation)*	.079 *** (2275)	.076 *** / n/a (2307)	.113 *** (3074)	.032 (3074)
	Income Deprivation Decile	.078 *** (2275)	.083 *** / n/a (2307)	.125 *** (3074)	.039 * (3074)
	Employment Deprivation Decile	.063 ** (2275)	.073 *** / n/a (2307)	.109 *** (3074)	.008 (3074)
	Health Disability Decile	.055 ** (2275)	.048 * / n/a (2307)	.098 *** (3074)	.016 (3074)
	Education Deprivation Decile	.056 ** (2275)	.064 ** / n/a (2307)	.083 *** (3074)	-.019 (3074)
	Housing and Services Deprivation Decile	.046 * (2275)	.035 / n/a (2307)	-.024 (.176)	.059 *** (3074)
	Crime Deprivation Decile	.061 ** (2275)	.040 / n/a (2307)	.100 *** (3074)	.062 *** (3074)
	Living Environment Decile	.049 * (2275)	.036 / n/a (2307)	.066 *** (3074)	.042 * (3074)
UKCAT measures	UKCAT questions skipped/missed	.000 (3432)	-.015 / -.041 (2764 / 769)	-.310 *** (4811)	-.005 (4811)
	UKCAT percentile day of taking test	-.092 *** (3432)	-.059 ** / -.018 (2764 / 769)	-.058 *** (4811)	-.090 *** (4811)
	UKCAT allowed extra time	.007 (3432)	.014 / -.012 (2764 / 769)	.030 * (4811)	.004 (4811)
	UKCAT School experience of test	-.029 (3295)	.038 / -.174 *** (2630 / 754)	-.033 * (4022)	-.033 * (4022)

Table 6. Simple Pearson correlations of the four UKCAT subscales with each other and with the total score. N=4811 in all cases, except with Educational Attainment, where N=3432, and with A-levels and Highers where N is 2764 and 769. . *** p<.001; **: p<.01; * p<.05. Correlations in brackets indicate the correlations between the subscales disattenuated for reliability.

	Abstract Reasoning	Decision Analysis	Quantitative Reasoning	Verbal Reasoning
Abstract Reasoning	1	.196*** (.284)	.190*** (.249)	.114*** (.155)
Decision Analysis	.196*** (.284)	1	.156*** (.243)	.146*** (.236)
Quantitative Reasoning	.190*** (.249)	.156*** (.243)	1	.213*** (.311)
Verbal Reasoning	.114*** (.155)	.146*** (.236)	.213*** (.311)	1
Total UKCAT score	.604***	.655***	.583***	.591***
Educational Attainment	.144***	.131***	.133***	.087***
3 best A-levels/ 5 best Highers	.123***/ .083 *	.121***/ .129***	.127***/ .202***	.062**/ .070

Table 7. Simple Pearson correlations of the three medical school outcome measures with the four UKCAT sub-scores, the UKCAT total score, and the measure of educational achievement. **Key**: * P<.05; ** P<.01; *** P<.001. Correlations with p<.05 are also shown in bold.

	OverallMark	SkillsMark	TheoryMark
Abstract Reasoning	**.080** *** **(4811)**	**.053** ** **(3184)**	**.052** * **(2075)**
Decision Analysis	**.090** *** **(4811)**	**.056** *** **(3184)**	**.077** *** **(2075)**
Quantitative Reasoning	**.076** *** **(4811)**	**.044** * **(3184)**	**.079** *** **(2075)**
Verbal Reasoning	**.115** *** **(4811)**	.028 (3184)	**.177** *** **(2075)**
Total UKCAT score	**.148** *** **(4811)**	**.075** *** **(3184)**	**.160** *** **(2075)**
Educational Attainment	**.362** *** **(3432)**	**.210** *** **(2240)**	**.351** *** **(1407)**
3 best A-levels/ 5 best Highers	**.177** *** / .003 **(2764** / 769**)**	**.096** *** / .027 **(2000** / 298**)**	**.248** *** / .074 **(1250** / 199**)**

Table 8: Comparison of the four outcome groups in relation to the various continuous measures. Values in cells are mean (SD, N).

	Fail (a)	Repeat 1st year (b)	Passed after resits (c)	Passed all first time (d)	ANOVA (r) Linear trend F(1,n)	ANOVA (r) Nonlinear F(2,n)	Levene test	Homogenous subsets (s)
Overall Mark (p, q); All cases	**-2.644** (1.28, 96)	**-1.924** (.99, 94)	**-1.110** (.79, 565)	**.235** (.80, 4056)	2843.8 P<.001	36.9 P<.001	P<.001	a,b,c,d
Overall Mark (p, q); Cases only with explicit overall marks	**-1.877** (.74, 63)	**-1.726** (.90, 76)	**-1.117** (.82, 516)	**.223** (.82, 3855)	1755.6 P<.001	80.2 P<.001	NS	ab, c, d
Theory Mark (p)	**-1.258** (.44, 29)	**-1.322** (.82, 29)	**-.654** (.76, 294)	**.250** (.68, 1723)	619.9 P<.001	23.1 P<.001	NS	ab, c, d
Skills Mark (p)	**-1.079** (.84,40)	**-.891** (.87, 438)	**-.616** (.90, 438)	**.214** (.71, 2655)	602.8 P<.001	32.0 P<.001	P<.001	ab, bc, d
UKCAT total	**2492** (192, 96)	**2457** (230,94)	**2486** (205, 565)	**2544** (205,4056)	43.7 P<.001	6.8 P=.001	NS	abc, ad
zUKCAT (p)	**-.121** (.95,96)	**-.312** (1.01, 94)	**-.186** (.99, 565)	**.036** (.99, 4056)	25.3 P<.001	5.3 P=.005	NS	abc, ad
zUKCAT abstract reasoning (p)	**-.163** (.94, 96)	**-.224** (1.02, 94)	**-.096** (.97, 565)	**.022** (1.00, 4056)	13.1 P<.001	0.75 NS	NS	abcd
zUKCAT Decision Analysis (p)	**-.064** (1.08, 96)	**-.302** (.99, 94)	**-.129** (.98, 565)	**.026** (.995, 4056)	14.01 P<.001	3.70 P=.025	NS	abc, ad
zUKCAT Quantitative Reasoning (p)	**.045** (.97, 96)	**-.110** (1.03, 94)	**-.116** (1.06, 565)	**.018** (.99, 4056)	3.38 NS	3.38 P=.034	NS	abcd
zUKCAT Verbal Reasoning (p)	**-.087** (.97, 96)	**-.136** (1.16, 94)	**-.127** (.96, 565)	**.023** (1.00, 4056)	9.56 P=.002	2.13 NS	NS	abcd
zEducational Attainment (p)	**-.441** (.942, 65)	**-.653** (1.06, 60)	**-.563** (1.14, 414)	**.104** (.94, 2893)	156.1 P<.001	29.5 P<.001	P<.001	abc, d
Three best A-levels	**28.89** (2.82, 56)	**28.28** (1.58, 49)	**29.06** (1.36, 333)	**29.39** (1.22)	44.8 P<.001	6.47 P=.002	P<.001	a,bc,d
Five best Highers	**48.71** (1.57, 17)	**47.50** (2.58, 16)	**48.70** (2.46, 88)	**48.99** (2.40, 648)	3.8 NS	1.66 NS	NS	abcd

Notes:

p. values are standardised within schools (Note: this is not the case for the raw, UKCAT total mark)

q. In a small proportion of cases, as described in the text, the overall mark is based on a normal score derived from the four point categorical scale. For comparability with other analyses, the first row includes these cases. However the second row analyses only cases where an overall mark was explicitly provided.

r. The denominator df, n, can be calculated as N-4, N is the total number of cases (provided in individual cells)

s. If values are together then they are not significantly different from one another and form a homogenous subset with p>.05 using the Ryan-Einot-Gabriel-Welsch range test. As an example, "abc, ad" means that groups a, b and c (Fail, repeat 1st year and passed after resits) do not differ from one another; likewise groups a and d (Fail, Passed all first time) do not differ from one another. Group d (Passed all first time) is significantly different from Repeat 1st year and Passed after resits. "a,b,c,d" indicates each differs from each of the other three groups, and "abcd" indicates no significant post hoc differences.

Table 9: Comparison of those leaving medical school for academic reasons with those leaving for non-academic reasons. Values in cells are mean (SD, N).

	Academic	Non-academic	Levene test for heterogeneity	Significance of difference in means (t-test)
Overall Mark	**-2.13** (.95, 55)	**-2.57** (1.76, 49)	**P<.001**	P=.107
Overall Mark (p, q); Cases only with explicit overall marks[1]	**-2.13** (.91, 32)	**-3.12** (1.76, 20)	**P=.002**	**P=.011**
Theory Mark (p)	**-1.31** (.42, 25)	**-1.64** (1.01, 5)	**P=.001**	P=.232
Skills Mark (p)	**-1.05** (.80, 41)	**-1.42** (1.11, 11)	P=.063	P=.227
UKCAT total	**2493.8** (181.5, 55)	**2504.5** (232.1, 49)	**P=.009**	P=.797
zUKCAT (p)	**-.194** (.90, 55)	**.018** (1.02, 49)	P=.399	P=.261
zUKCAT abstract reasoning (p)	**-.196** (.92, 55)	**-.145** (1.03, 49)	P=.358	P=.802
zUKCAT Decision Analysis (p)	**-.126** (1.06, 55)	**-.021** (1.17, 49)	P=.633	P=.633
zUKCAT Quantitative Reasoning (p)	**-.002** (.90, 55)	**-.194** (.99, 49)	P=.616	P=.292
zUKCAT Verbal Reasoning (p)	**-.078** (1.01, 55)	**.016** (.94, 49)	P=.825	P=.627
zEducational Attainment (p)	**-.631** (.90, 39)	**-.221** (.88, 35)	P=.828	**P=.051**
Three best A-levels	**29.09** (1.22, 35)	**28.81** (4.26, 22)	P=.199	P=.727
Five best Highers	**48.86** (1.07, 7)	**48.82** (2.01, 17)	P=.271	P=.967

[1] In a small proportion of cases, as described in the text, the overall mark is based on a normal score derived from the four point categorical scale. For comparability with other analyses, the first row includes these cases. However the second row analyses only cases where an overall mark was explicitly provided.

Table 10. Calculation of validity coefficients for A-levels and UKCAT, corrected for restriction of range and unreliability of selection and performance measures, using the method of Hunter at al (2006) (see their table 2, p.603). Note that in the terminology of Hunter et al, X = Selection Measure, T = True score component of selection measure, Y = the actual performance measure, and P = the true score component of the performance measure, with a=applicants and i=incumbents (i.e entrants).

Selection measure (X)		Three best A-levels		UKCAT total
Unrestricted group		Medical school applicants	All UCAS applicants	Medical school applicants
Performance measure (Y)		First year performance	First year performance	First year performance
Input measures	Formula[2]			
Correlation of selection measure and medical school performance in the restricted group of medical school entrants	ρ_{XY_i}	.177	.177	.148
Reliability of selection measure in the unrestricted applicant population	ρ_{XX_a}	.928	.969	.865
Reliability of performance measure in the restricted group of medical school entrants	ρ_{YY_i}	.840	.840	.840
Range restriction on selection measure (SD(entrants)/SD(applicants))	u_X	.335	.219	.791
Output measures				
Reliability of selection measure in the restricted population	ρ_{XX_i}	.359	.359	.784
Operational validity: Correlation of raw selection measure and disattenuated performance measure in the unrestricted population	ρ_{XP_a}	.822	.916	.222
Construct validity: Correlation of disattenuated selection measure and disattenuated performance measure in the unrestricted population	P_{TPa}	**.854**	**.930**	**.239**

[2] Notation used by Hunter et al, 2006

Figure 1: Histograms of *OverallMark*, *TheoryMark* and *SkillsMark*, and a scattergram of *SkillsMark* in relation to *TheoryMark*. The red line is the linear regression, and the green line a lowess curve.

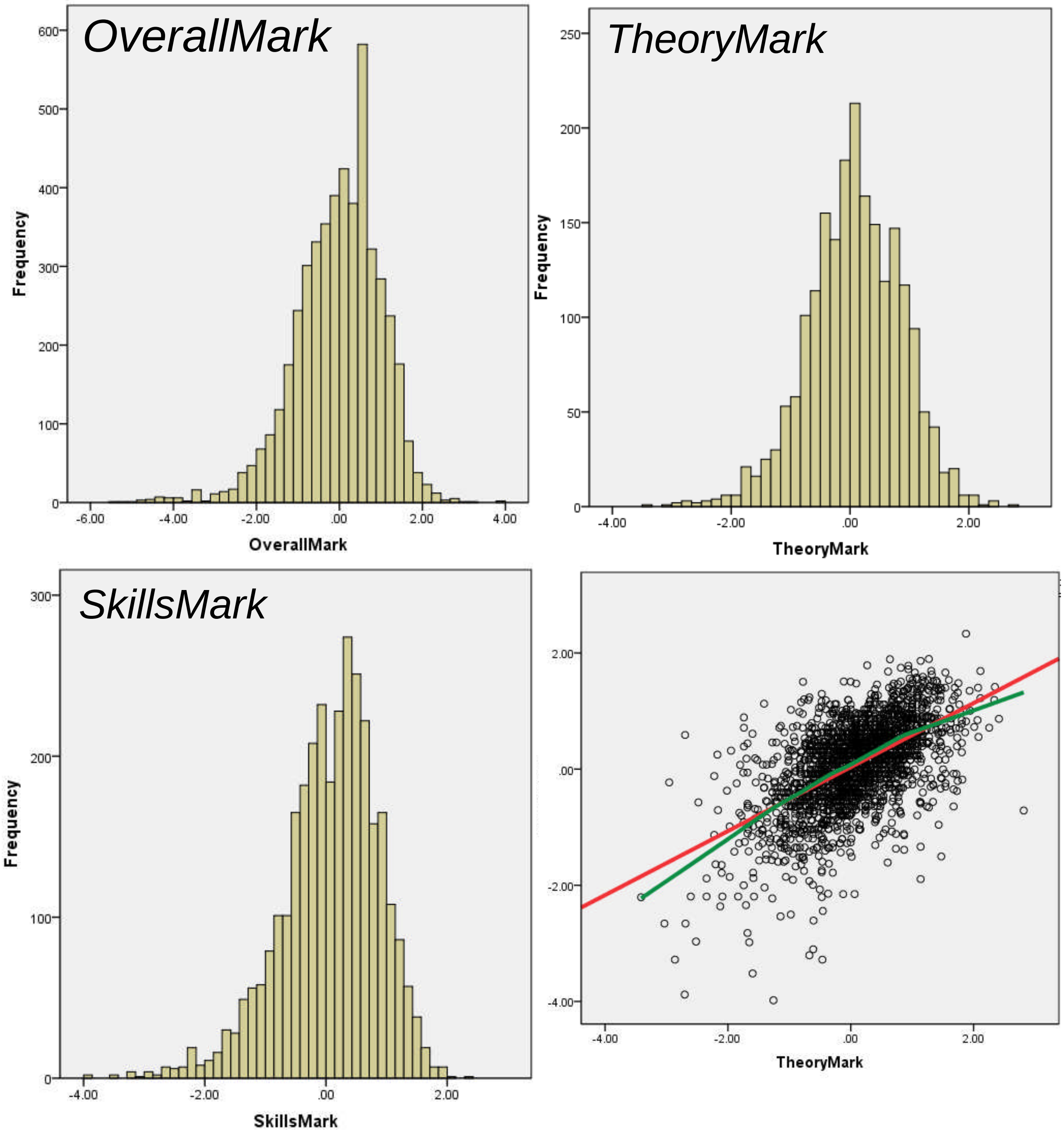


Figure 2: a) Distribution of *zEducationalAttainmen*t, a composite score based on A-levels, AS-levels and GCSEs and b) its prediction of *Overall Mark* by *zEducationalAttainment*. The red line is a standard linear regression, and the green line is a lowess curve.

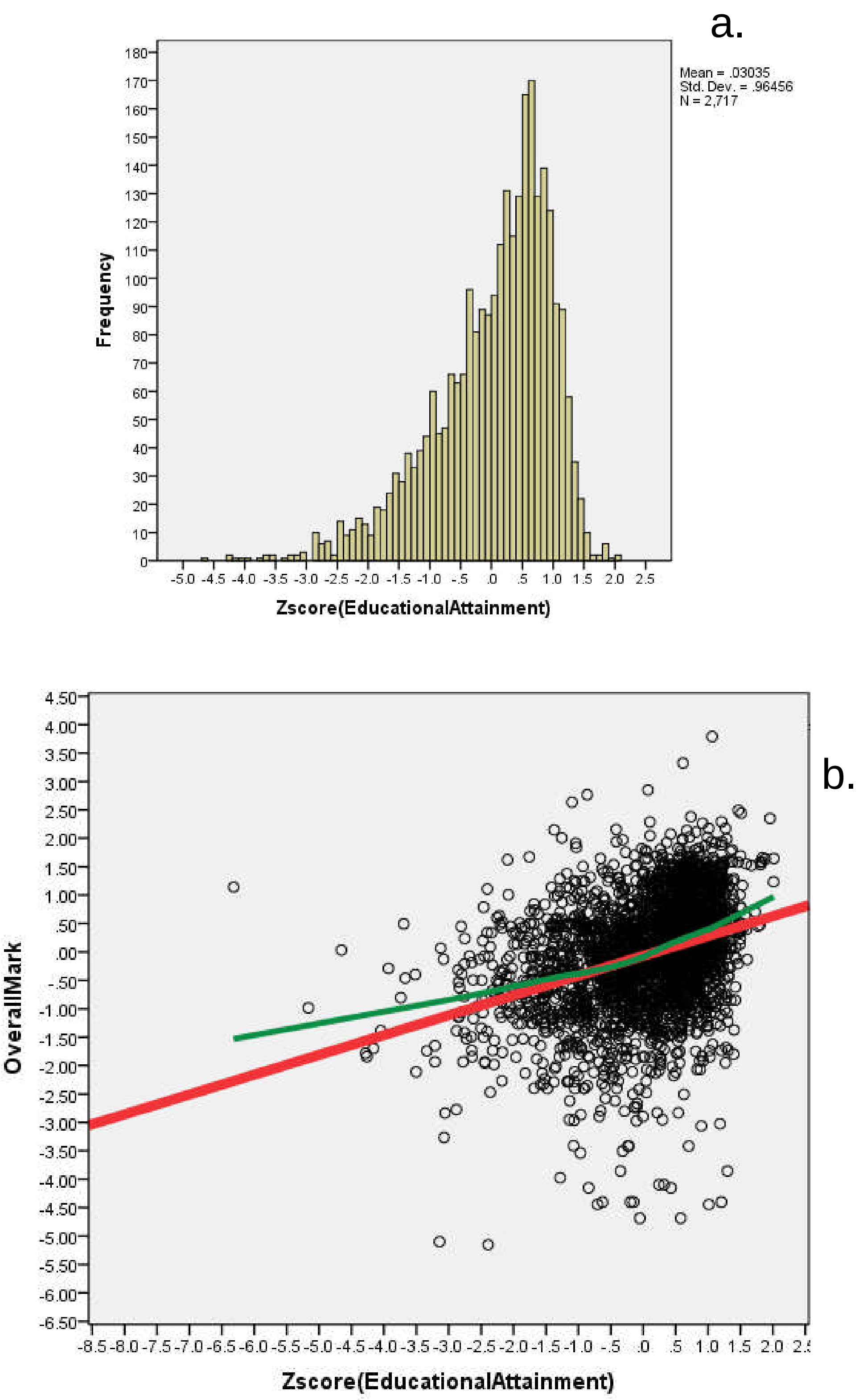


Figure 3: a) Distribution of *zEducationalAttainmen*t, a composite score based on Scottish Highers, ‘HighersPlus’ and Advanced Highers; and b) its prediction of *Overall Mark* by *zEducationalAttainment*. The red line is a standard linear regression, and the green line is a lowess curve. The x and y scales of the scattergram are identical to those of figure 2, and it can be seen that the slope is greater for Highers than for A-levels.

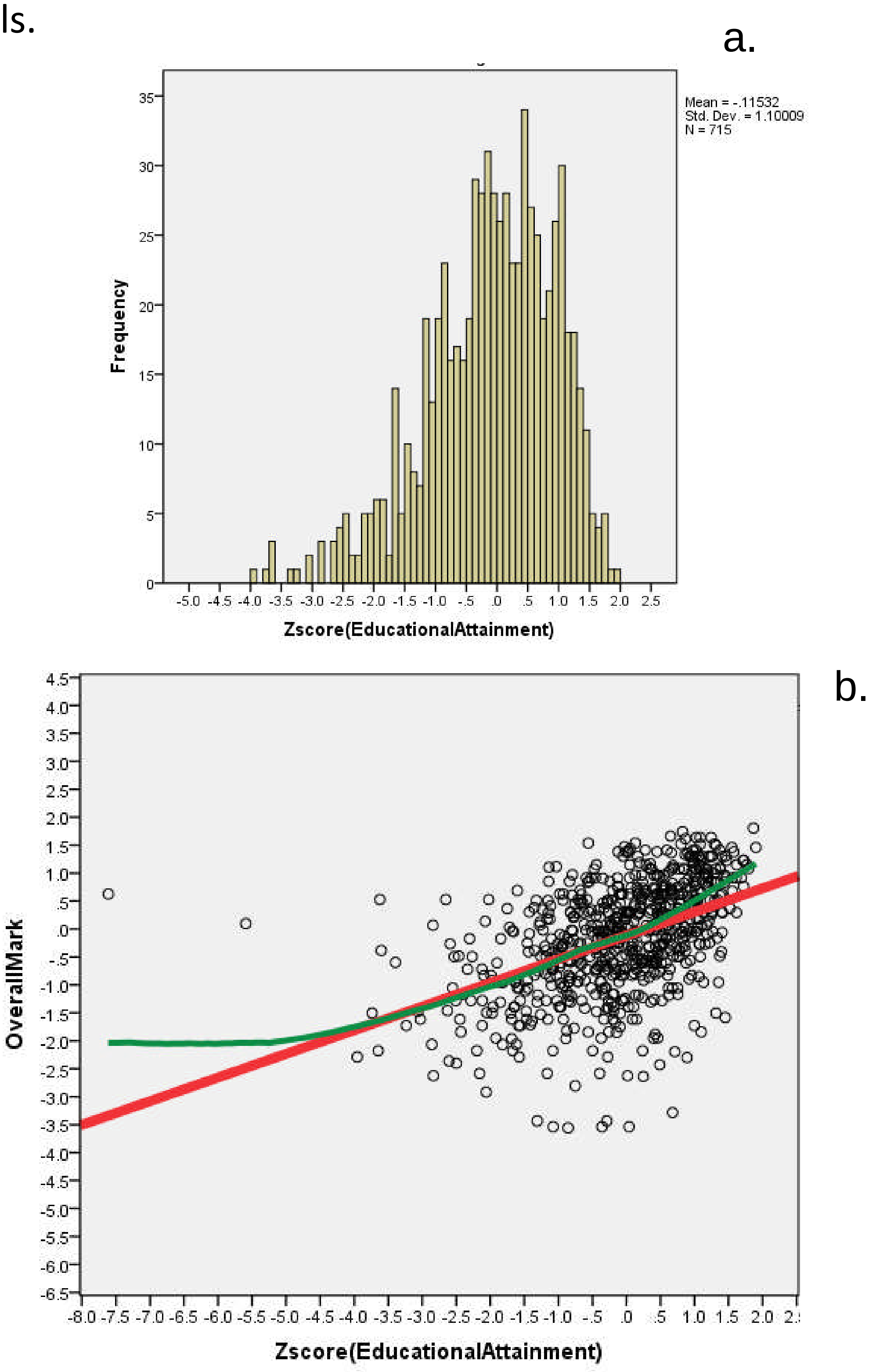


Figure 4: Distribution of *Overall Mark* in relation to *zEducationalAttainment*, separately for White (blue points and line) and non-White (green points and line) students. The fitted lines are lowess curves.

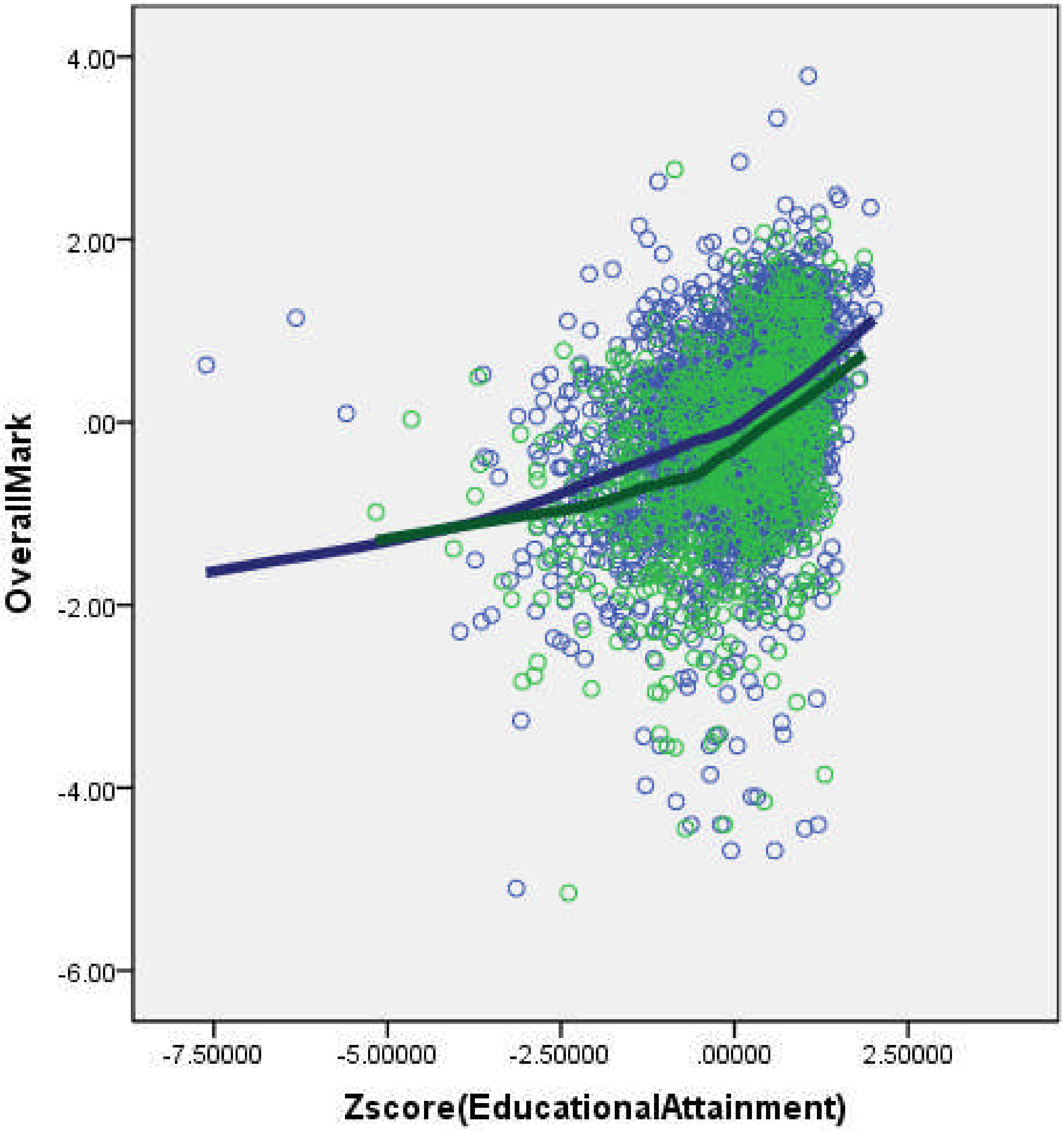


Figure 5: Scattergram of performance on UKCAT total score in relation to date of taking of UKCAT in days after the test opened for the testing season (and the seasons are of different lengths in different years). The three cohorts are shown separately (2007: blue; 2008: green; 2009: red). Fitted lines are lowess curves.

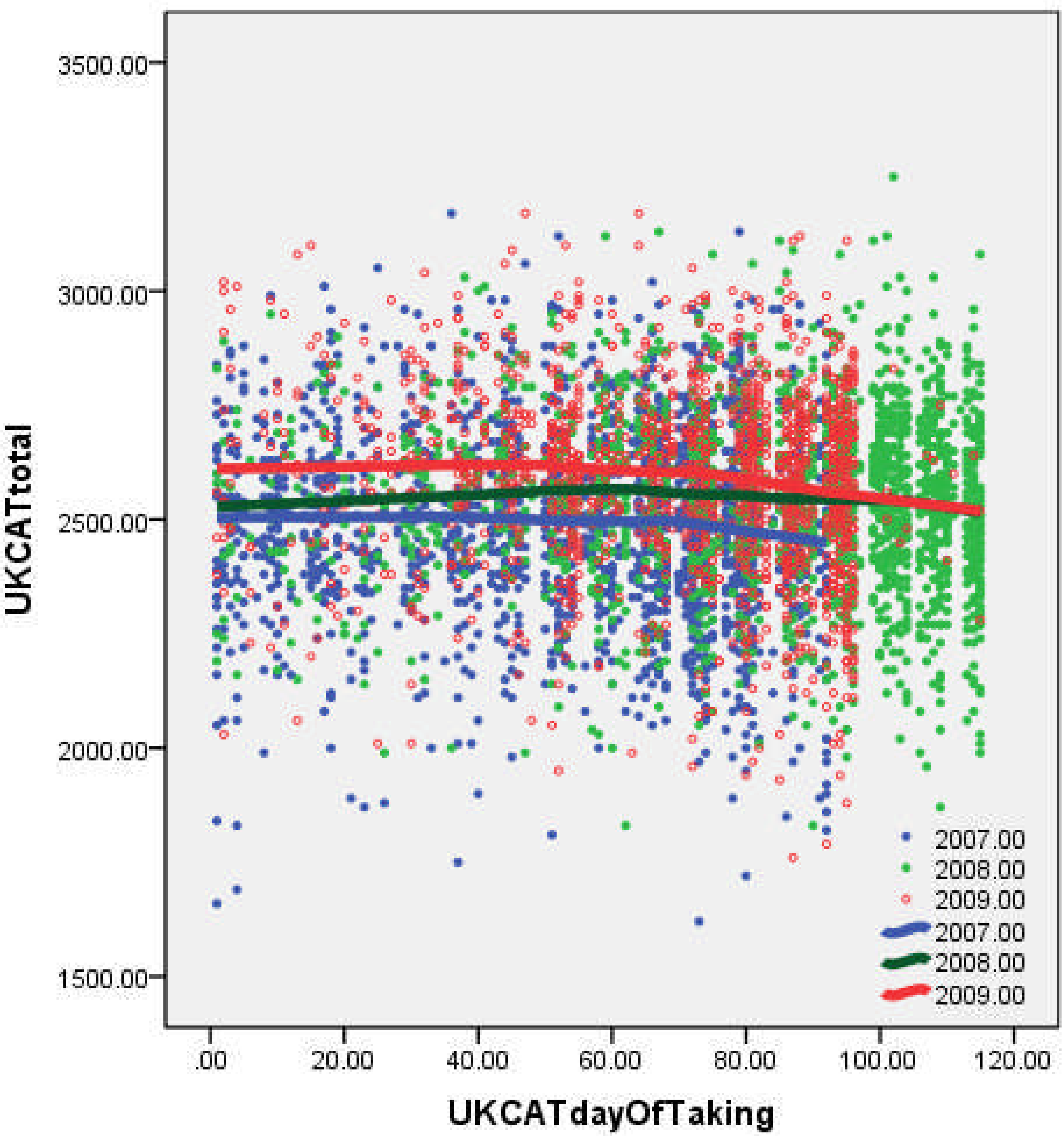


Figure 6: Scattergram showing relationship between overall mark at medical school, and UKCAT score (standardised within medical schools). Mature students (green) and non-mature students (blue) are shown separeately, along with fitted linear regression functions. The crossing of the two lines is at about 2.5 standard deviations below the mean, so that at almost all candidate ability levels, mature students outperform non-mature students, with a steeper slope for mature students.

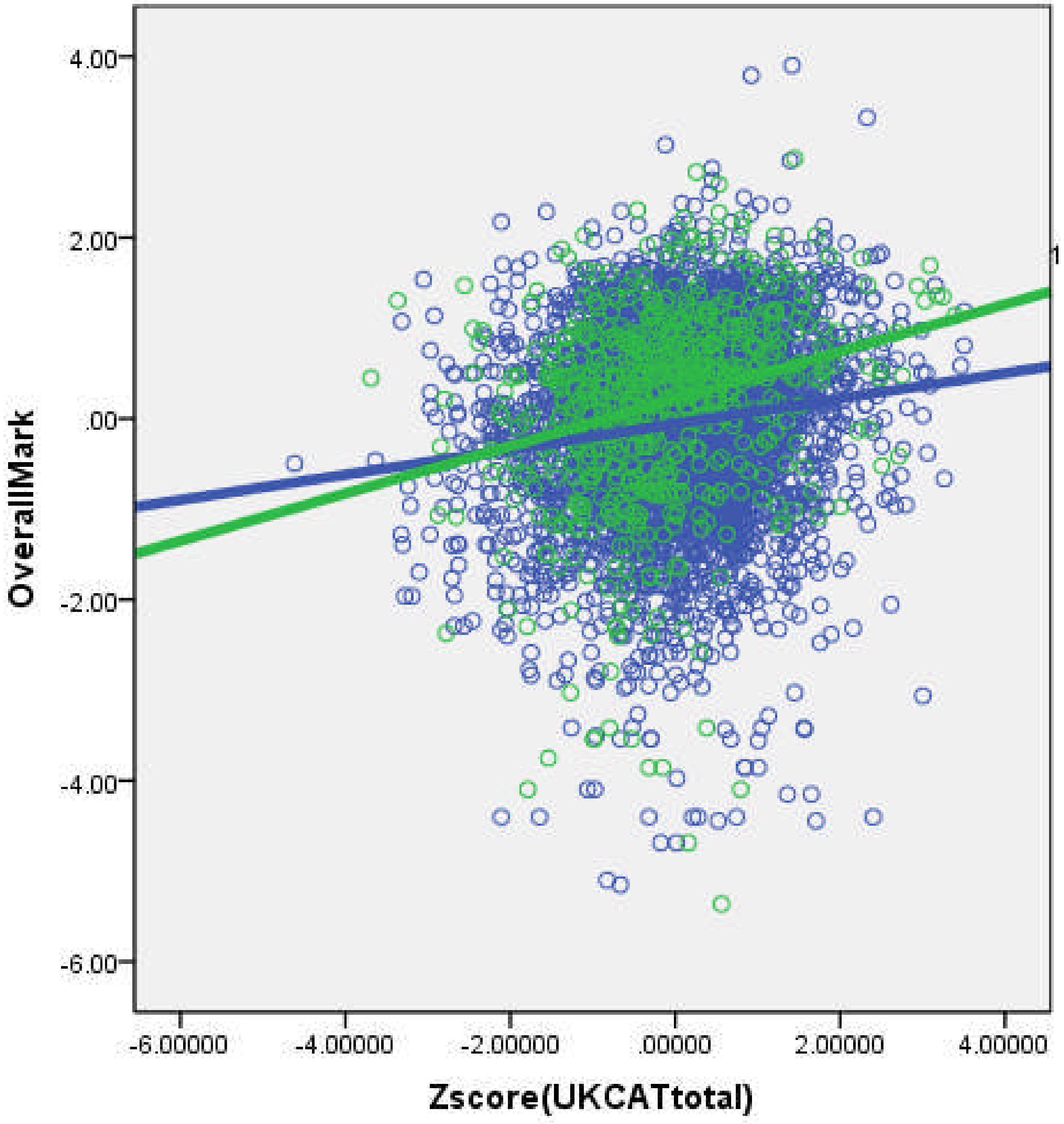


Figure 7: Graph indicating the relationship between first year examination performance and entrance grades based on three best A-levels. Examination performance has been divided into four groups: students who pass all exams first time (purple); students who progress to the second year, with or without first year resits (green); students who have to repeat the first year or leave the medical school (orange); or those who leave the medical school for academic or other reasons (red). Curves are calculated by extrapolation from logistic regression carried out on the entrants to medical school. As curves move away from the typical marks of entrants then they become less precise, but nevertheless give an indication of the likely effects were entrants to have lower A-level grades than is conventional at present in most schools. The 95% range of grades in medical students and medical school applicants is shown by the bars at the top of the figure. The overall range is for A-level grades of possible university applicants.

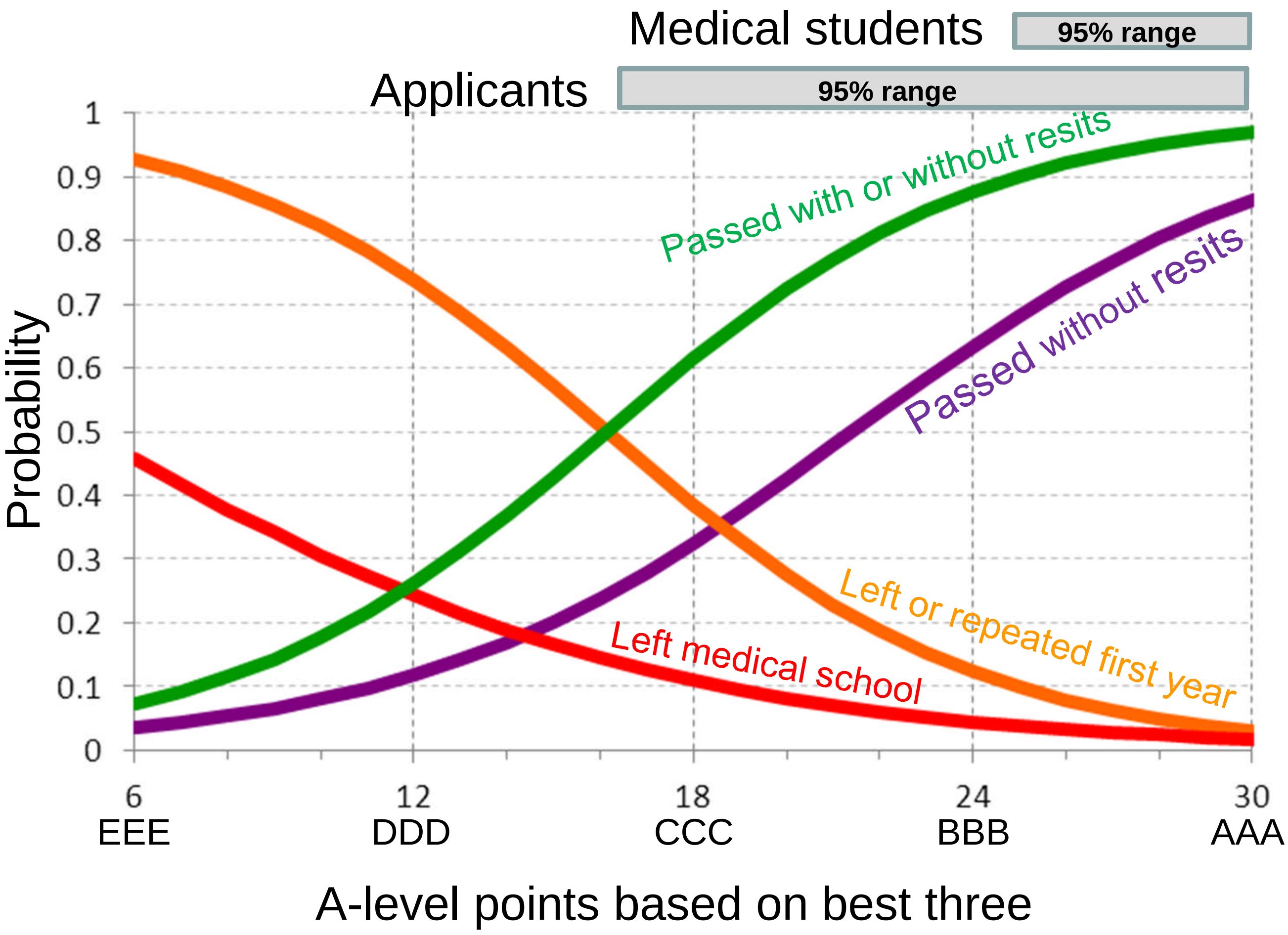
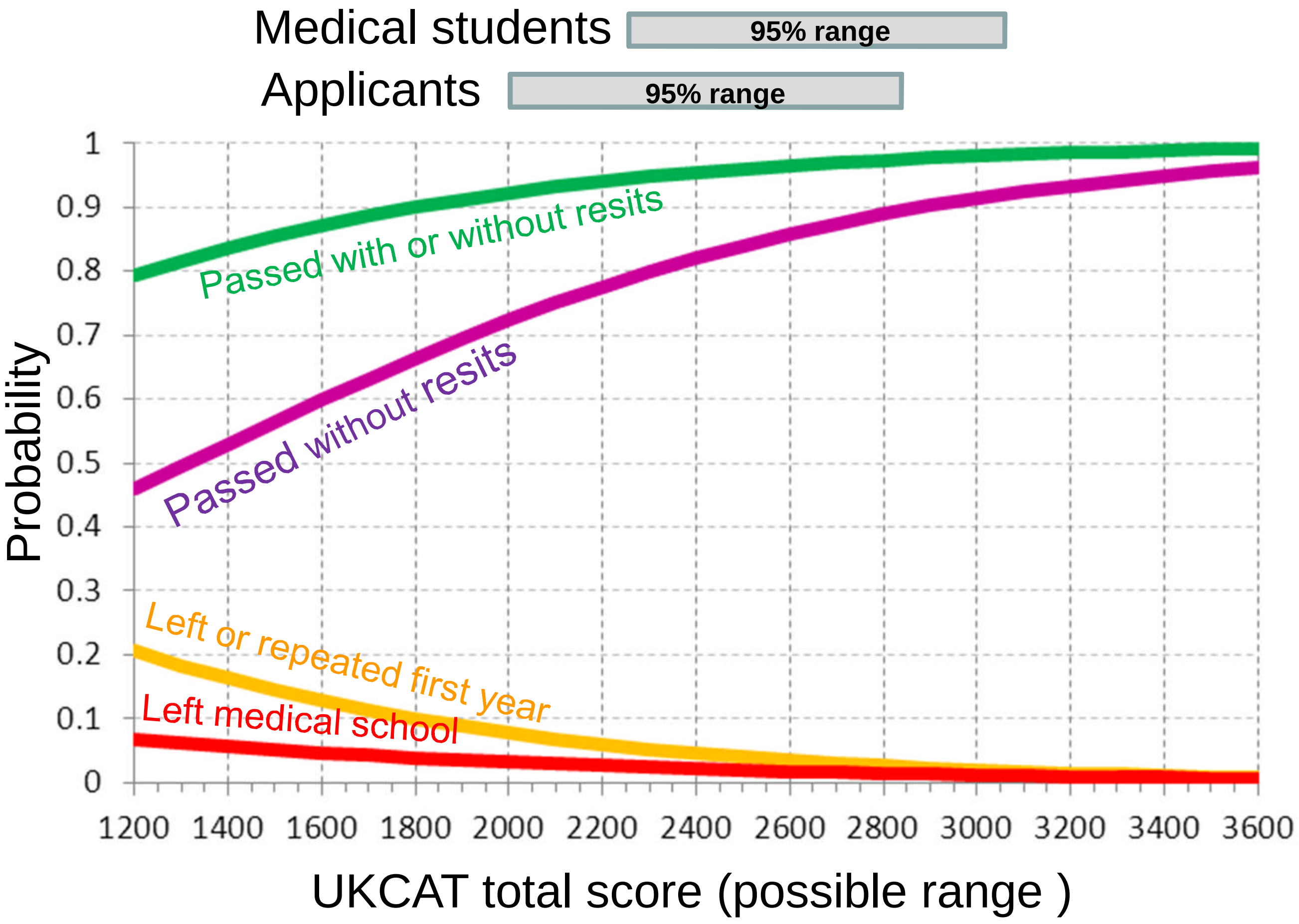


Figure 8: Graph indicating the relationship between first year examination performance and UKCAT total scores. Examination performance has been divided into four groups: students who pass all exams first time (purple); students who progress to the second year, with or without first year resits (green); students who have to repeat the first year or leave the medical school (orange); or those who leave the medical school for academic or other reasons (red). Curves are calculated by extrapolation from logistic regression carried out on the entrants to medical school. As curves move away from the typical marks of entrants then they become less precise, but nevertheless give an indication of the likely effects were entrants to have lower A-level grades than is conventional at present in most schools. The 95% range of grades in medical students and medical school applicants is shown by the bars at the top of the figure. The range of possible UKCAT marks is from 1200-3600.


Medical students
95% range
Applicants
95% range
Probability
Passed with or without resits
Passed without resits
Left or repeated first year
Left medical school
UKCAT total score (possible range)