Predictive validity of the Biomedical Admissions Test: An evaluation and case study

I. C. McManus1, Eamonn Ferguson2, Richard Wakeford3, David Powis4 & David James5

Abstract
There has been an increase in the use of pre-admission selection tests for medicine. Such tests need to show good psychometric properties. Here, we use a paper by Emery and Bell [2009]. The predictive validity of the Biomedical Admissions Test for pre-clinical examination performance. Med Educ 43:557–564 as a case study to evaluate and comment on the reporting of psychometric data in the field of medical student selection (and the comments apply to many papers in the field). We highlight pitfalls when reliability data are not presented, how simple zero-order associations can lead to inaccurate conclusions about the predictive validity of a test, and how biases need to be explored and reported. We show with BMAT that it is the knowledge part of the test which does all the predictive work. We show that without evidence of incremental validity it is difficult to assess the value of any selection tests for medicine.

Introduction
The use of cognitive ability tests to help select medical students is becoming more common in the UK, as well as elsewhere (Ferguson et al. 2002; Parry et al. 2006). Two factors have prompted this rise in use. First, the drive to increase the diversity of students studying medicine (Patterson & Ferguson 2007; Powis et al. 2007a; James et al. 2008, 2009), an argument which is based on the claim that traditional academic selection tools (e.g. A levels) are inherently biased, favouring female applicants from high social groups and disadvantaging good potential male candidates from lower SES groups (Powis et al. 2007a). The need for diversity has been argued on the basis that health care provision would be improved if the diversity of the local community were matched by that of physicians (Carrasquillo & Lee-Rey 2008; Saha et al. 2008). Second, it has been argued that traditional academic selection tools – especially A levels in the UK – as well as having possible social biases are unable to distinguish between the many able candidates applying to study medicine, due to a majority of would-be students achieving the maximum grades of AAA (McManus et al. 2005).

Based on such arguments, a variety of cognitive ability tests have been developed (McManus et al. 2005), such as Biomedical Admissions Test (BMAT), UK Clinical Aptitude Test (UKCAT) and Graduate Medical Schools Admissions Test (GAMSAT). Any such test should be reliable and valid. Reliability is a relatively straightforward concept (Downing 2004). However validity is more complex (Downing 2003; Lissitz 2009), and as Kane (2009) has put it, ‘Validity is simple; validation can be difficult’. Hutchinson et al. (2002) have provided a useful list of different ways of assessing both reliability and validity. Tests should also be free of biases for age, sex, ethnicity and socio-economic background (although that should not be interpreted as an absence in difference in mean performance between groups, which may well represent true group differences, but more properly should mean that there is no differential item functioning (Hambleton et al. 1991).

Practice points
• Cognitive tests used in the selection of medical students, which primarily assess fluid intelligence, or crystallised intelligence (e.g. knowledge tests based on taught material) should be reliable and valid, free from bias (e.g. gender and ethnicity) and show good psychometric properties.
• To highlight the importance of these properties to medical education and selection, this article explores, as a case study, the psychometric properties of BMAT, using the information presented in the study of Emery and Bell (2009).
• No published information has been provided on the reliability of BMAT, which provides challenges to the proper interpretation of the test.
• BMAT Section 2 (Scientific Knowledge and Applications), a measure of crystallised intelligence, has a medium correlation of 0.36 with year 1 examination results, and a weak correlation of 0.23 with results in year 2.
• BMAT Section 1 (Aptitude and Skills), a measure of fluid intelligence, has no incremental validity over Section 2, the partial correlations of 0.054 and 0.065 with year 1 and year 2 performance effectively being zero.
chapter 8) for individuals from particular backgrounds (i.e. the measurement properties of the test are invariant across groups), as has indeed been demonstrated for UKCAT (UKCAT 2008a, 2008b, 2009).

A fundamental issue in evaluating any such test is the need to show that such measures are free from the inherent problems claimed for traditional academic selection measures.

A central claim for the newer measures such as BMAT and UKCAT is that they tap basic potential – which is termed ‘aptitude’ by most test developers. A distinction can be made between cognitive ability tests that tap raw cognitive abilities (termed ‘fluid intelligence’) and those which reflect knowledge and learning acquired through education (termed ‘crystallised intelligence’; Blair 2006). The latter inevitably measures knowledge and skills acquired at school and these represent up to 93% of all intelligence tests (Blair 2006). It is this type of intelligence or cognitive ability that medical selection tests will mainly be tapping. In that respect they are likely to show similar social correlations as do A level assessments.

Any cognitive selection test should, at the very least, show reliability and validity. Reliability shows that the test scores are equivalent or stable over time and that the questions are measuring the same content (internal consistency). Reliability is a precursor for validity and in this regard the measures should have not only content and construct validity but also predictive and incremental validity – that is they are able to predict future performance in the specified domain. Incremental validity also requires them to predict better than traditional measures, which in the UK would mean they predict medical school performance above and beyond the predictive ability of A level scores. Unfortunately, data reported in many journals do not contain all the necessary psychometric detail to evaluate issues, such as predictive validity (Ferguson et al. 2002).

Rather than review, a wider literature on reliability and validity of medical selection test in general, we here focus on a single paper as a case study to highlight the potential pitfalls in interpreting evidence if complete evidence is not published. We felt that a case study approach makes the arguments clearer as the reader can refer to a single concrete example. We chose the Emery and Bell (2009) paper as it is a very recent publication in the area, and so represents contemporary reporting of such data, in this case on BMAT, and is not open to the criticism that the case study is old and out of date. We use the data published in the Emery and Bell paper to highlight some of the problems when full psychometric data are not presented. The results reported in the paper are typical of many papers published in medical selection and should be seen as a vehicle to highlight salient issues that pertain to the wider literature, and to show the type of interpretative information that is required. In particular, we will show how the simple reporting of zero-order correlations can lead to misleading results that over-estimate the predictive validity of the test. We base our evaluation around the issues that have been raised previously by McManus et al. (2005).

**The BMAT**

The BMAT is in two sections. Section 1 is the ‘Aptitude and Skill’ section and it assesses problem solving, argument comprehension, data and graphical interpretation. Section 2 assesses ‘Scientific Knowledge and Application’, specifically knowledge in biology, chemistry, physics and maths to the level of non-specialist schools in the UK up to age 16. Both sections are assessed using MCQs. Section 2 is explicitly an index of crystallised intelligence but, from the description, so in part is Section 1. Marks on the two sections are reported as having a correlation coefficient of about 0.40, but they are nevertheless treated as separate assessments (Emery & Bell 2009).

**BMAT: A case study and evaluation of Emery and Bell (2009)**

**Differentiating acceptances from rejections**

Information on basic descriptive statistics (means, standard deviations (SDs), etc.) should be provided for selected and non-selected candidates as they aid interpretation of predictive validity. No information in Emery and Bell (2009) is provided on the means and SDs of the scores of accepted and rejected applicants. It is possible to infer from an earlier report (Emery & Bell 2006) that for those applying in 2003, the entrants’ mean score on Section 2 of the test was about 5.81 (SD 0.87), compared with the rejects’ mean of about 5.03 (0.78), a difference of 0.78 score points, which is about one SD (a difference that in part reflects the use of the test itself in selection).

**Reliability**

Any paper describing a cognitive ability tests should provide reliability data. UKCAT (2009) has reported internal reliabilities for its subtests which have lengths of 26, 40, 44 and 65 items of 0.58, 0.62, 0.65 and 0.78, respectively, and in addition its total score of 175 items has a reliability of 0.86, the latter value being acceptable for a high-stakes test. It is also clear, that to a first approximation, reliability is proportional to the square root of the number of items. Emery and Bell (2009) provide no reliability information for either Section 1 or Section 2, which have 27 and 35 items.

Reliability needs to be provided for a number of key interpretive reasons. Firstly, reliability is a pre-cursor for validity, as a test that is not reliability cannot be valid. Secondly, when looking at predictive validity it is useful to correct validity coefficients for lack of reliability as that allows an estimate of the disattenuated correlation between the predictor and outcome, had the test been perfectly reliable and as such gives an upper limit of validity. Such information is therefore, crucial when interpreting the size of the predictive validity coefficients. Emery and Bell (2009, p. 559) do indicate that Sections 1 and 2 are correlated at about 0.40, indicating some consistency.

**Combining data across years**

When there is a large data set it is useful to combine estimates to achieve a more reliable overall estimate than report numerous associations, without any correction for chance.
Validity of the BMAT

Table 1. Random effects meta-analyses of Table 2 from Emery and Bell (2009).

<table>
<thead>
<tr>
<th>BMAT section</th>
<th>Year</th>
<th>Course component</th>
<th>Meta-analyses for individual course components</th>
<th>Meta-analyses for total marks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 1</td>
<td>Homeostasis</td>
<td>$r = 0.17^*; Q = 2.05\text{ns}$</td>
<td>$r = 0.19^*; Q = 1.63\text{ns}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Molecules in medical science</td>
<td>$r = 0.20^*; Q = 3.37\text{ns}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Functional architecture of the body</td>
<td>$r = 0.14^*; Q = 0.99\text{ns}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Year 2</td>
<td>Biology of disease</td>
<td>$r = 0.16^*; Q = 0.86\text{ns}$</td>
<td>$r = 0.15^*; Q = 2.20\text{ns}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Human reproduction</td>
<td>$r = 0.13^*; Q = 1.68\text{ns}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Neurobiology and human behaviour</td>
<td>$r = 0.14^*; Q = 5.78\text{ns}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mechanisms of drug action</td>
<td>$r = 0.20^*; Q = 0.99\text{ns}$</td>
<td></td>
</tr>
<tr>
<td>Section 2</td>
<td>Year 1</td>
<td>Homeostasis</td>
<td>$r = 0.37^<em>; Q = 10.24^</em>$</td>
<td>$r = 0.36^*; Q = 4.21\text{ns}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Molecules in medical science</td>
<td>$r = 0.36^*; Q = 4.21\text{ns}$</td>
<td>$r = 0.42^<em>; Q = 12.99^</em>$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Functional architecture of the body</td>
<td>$r = 0.25^*; Q = 5.28\text{ns}$</td>
<td>$r = 0.20^*; Q = 0.99\text{ns}$</td>
</tr>
<tr>
<td></td>
<td>Year 2</td>
<td>Biology of disease</td>
<td>$r = 0.33^*; Q = 5.80\text{ns}$</td>
<td>$r = 0.23^*; Q = 1.04\text{ns}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Human reproduction</td>
<td>$r = 0.24^*; Q = 4.40\text{ns}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Neurobiology and human behaviour</td>
<td>$r = 0.25^*; Q = 5.28\text{ns}$</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mechanisms of drug action</td>
<td>$r = 0.33^*; Q = 6.00\text{ns}$</td>
<td></td>
</tr>
</tbody>
</table>

For example, Emery and Bell (2009) report 90 simple zero-order correlations for the BMAT Sections and with a variety of exam performances in years 1 and 2 of the medical degree collected over a 4-year period. Rather than analyse the data separately for each year, data could be combined across the 4 years to predict each exam outcome, so that a more reliable estimate is available. We therefore undertook a series of random effects meta-analyses using estimates for each of the 4 years to predict each outcome from the data reported by Emery and Bell (2009). These resulted in estimates for each exam paper, and also for all the papers in each year (see Table 1; which is in the same format as Emery and Bell’s (2009) Table 2). The overall coefficient of predictive validity is $r$, averaged across all relevant correlations. $Q$ indicates the extent to which the coefficients contributing to the average coefficient are homogeneous across the 4 years, with a significant value indicating that they are heterogeneous. Three of the 16 $Q$ coefficients are statistically significant at $p < 0.05$. Using the conventional descriptions of Cohen (1992), the correlations with total of 0.19, 0.15 and 0.25 are ‘small’, and only the correlation of 0.36 can be described as ‘medium’.

Table 1 shows clearly that the effect sizes for Section 2 (Scientific knowledge and Applications) are larger than for Section 1 (Aptitude and Skills). Considering the effect sizes in the Total column, the average correlation of 0.36 for Section 2 for year 1 is significantly larger than the average correlation of 0.19 for Section 1 for year 1 ($t(999) = 5.3, p < 0.0001$; the assumption is also made that Sections 1 and 2 are correlated 0.40, as Emery and Bell stated). Likewise, for year 2 the average correlation of 0.23 for Section 2 is significantly larger than the average correlation of 0.15 for Section 1 ($t(999) = 2.37, p = 0.017$). These results show clearly that Section 2 (Scientific Knowledge and Applications) is a stronger predictor of academic success than Section 1 (Aptitudes and skills).

Differential predictive power

Different components of a test will be correlated with each other and are also likely to have differential predictive power. That is, each aspect of a test will not carry equal weight when predicting a future event (e.g. exam performance). Thus Sections 1 and 2 may have differential predictive power.

The differences between Sections 1 and 2 were therefore, further explored by examining the partial correlations of each section with total academic performance in years 1 and 2. Considering firstly Section 2, and again assuming a correlation between sections of 0.40, the correlation with total year 1 exam score remains significant after taking into account the effect of Section 1 ($r = 0.316, p < 0.0001$); this correlation remains ‘medium’ using the criteria of Cohen (1992). However, the correlation of Section 1 with total year 1 exam score after taking Section 2 into account is non-significant ($r = 0.054, p = 0.089$). The correlation of Section 2 with the total year 2 score remains significant after partialling out Section 1 effects ($r = 0.19, p < 0.0001$), although the effect is ‘small’ using Cohen’s criterion, whereas the partial correlation of Section 1 with total year 2 score is barely significant and has an effect size that does not even reach Cohen’s criterion for ‘small’ ($r = 0.065, p = 0.040$). Such analyses show clearly that it is the knowledge component of the test (Section 2), rather than the aptitude component (Section 1) that is accounting for the predictive power of the test.

In summary, although Section 2, which is essentially a test of crystallised intelligence, has some predictive power (medium for Year 1 and small for Year 2), Section 1, which is largely a test likely to reflect fluid intelligence, appears to have only small predictive ability in its own right once Section 2 has been taken into account.

The above analyses show clearly that relying on simple zero-order correlations leads to a less than accurate picture of the predictive power of the test and also hides the fact that one component of the test is more important than the other. Future studies should probably use mixtures of structural modelling and multilevel modelling to fully interrogate their data (Ferguson et al. 2003).

Role of A levels and previous academic ability

It is implied by Emery and Bell (2009) that there is no variance in A-level grades, almost all entrants having three As, and hence that there can be no predictive validity of A-levels, and hence no need to assess incremental validity of BMAT over and above A-levels. It may be true that almost all entrants have three A-grades at A-level, but that is not necessarily the same
as there being no variance. The high scientific achievement at A level also makes it particularly surprising that Section 2, which the developers say is designed to measure science ability at National Curriculum Key Stage 4 (i.e. GCSE), can have predictive validity. However, even with no variance in grades in the best three A-levels, other individual variance will still be present. For example, it is of interest to know how BMAT (and university performance) correlates: (i) with number of A-levels (many candidates take 4 or more A-levels); (ii) with the number of science A-levels and (iii) with individual grades in Biology, Chemistry and Maths. A possible explanation for how Section 2 manages to be predictive is that because it assesses four different sciences (Maths, Physics, Chemistry and Biology), then those with A-levels in all four subjects will perform better both on Section 2 and their university examinations. Section 2 is, in other words, a surrogate or indicator, for A-levels, and hence would likely suffer from all of the perceived social deficits of A-levels. Further data and analyses are needed to evaluate. Finally, it might also be the case that BMAT, especially Section 2, or any other similar selection test, correlates with GCSE results, even for those with three A Levels (particularly in the case of the BMAT as this is intended to be at the level of GCSE science). The key issue here is whether scores on the new cognitive ability test show incremental predictive validity over GCSE performance, or are merely a surrogate for GCSE performance (and if there is no incremental validity over GCSE scores then GCSE scores can be used more cheaply, more efficiently and probably more reliably). In psychometric terms, the key issue here is that incremental validity is central to judging whether a test offers added predictive value, and incremental validity should always be reported.

Predicting degree class

There are a number of ways for examining predictive validity, with different outcome variables. Overall percentage scores are a continuous variable, but may be skewed, or degree class, which can be specified as 1st versus the rest and logistic regression used or 1st versus 2i versus 2ii versus 3rd etc and multi-nominal or ordinal regression used. There are relative merits of each – and it is often useful to explore if the same predictors arise with the different approaches. Emery and Bell (2009) chose to use logistic regression to predict a first class mark in the first or second year examinations (although it must be said that achieving a First is an unusual criterion given that medical school assessments, as with other assessments of professional ability, conventionally assess competence, rather than outstanding achievement). A problem with the Emery and Bell logistic regression is that odds ratios for each year cannot be compared, since the range of marks in each differs (hence an odds ratio of 1.21 for an increase in one mark in year 1 for MVAT 2000 where marks go from 12 to 38 bears no relation to an odds ratio of 1.7 for BMAT 2003, when marks go from 3.8 to 8.7). Providing standardised measures (e.g. odds ratio for a one SD increase in test score) would have allowed proper comparison. Another presentation and interpretive issue concerns the use of confidence intervals. This is illustrated by the graphs provided by Emery and Bell (2009). For example, Year 1 BMAT 2003 Section 2, the graphed line suggests that an entrant with a score of 8.7 (at the right hand end of the scale) would apparently have a 13.5 times higher chance of attaining a first compared with an entrant with a score of 3.8 at the left hand end of the scale. While statistically the estimate is correct, in reality it is difficult to interpret, for, [A] Typical BMAT candidate will score around 5.0... [The best candidates will score around 6.0, and a few exceptional candidates will score higher than 7.0. The upper mark of 8.7 used for the odds ratio calculation was actually the highest result achieved by any BMAT candidate in 2003. If confidence intervals are presented for the odds ratio of 13.5 then the limited predictive ability would have been immediately apparent. Likewise, confidence intervals should be plotted for the regression lines (Huff 1954).

A further problem of interpretation concerns the 'headline figure' of an odds ratio of 1.7 for regressing BMAT 2003 Section 2 marks on Year 1 exams, which sounds very impressive. However, looked at from the perspective that in that year the rejected candidates had a mean score only about 0.78 marks less than the mean scores of the accepted candidates (based on our best estimate from other published figures), it can be calculated that the accepted candidates were only about 1.5 times more likely to get a first than would have been the rejects, had they been admitted. That figure does not of course take into account other predictors of performance (A-level mix, GCSEs, school type, gender or whatever), so that the true odds ratio is likely to be substantially lower, to such an extent that it is not clear if such an odds ratio has any practical predictive value at all.

The above comments are designed to highlight some of the potential pitfalls, both in presenting odds ratios and in their interpretation, that can be easily avoided by careful plotting and presentation of confidence intervals, means and SDs.

Equality and social bias

The above analyses and arguments indicate that the BMAT's predictive power resides in Section 2, the knowledge component, which is close in structure to a standard academic test akin to A levels (indeed is explicitly based on GCSE level knowledge). Given that standard academic assessments are, for a host of reasons, influenced by sex, ethnicity, socio-economic background and so on (Powis et al. 2007b; McManus et al. 2008), predictive test that are based on knowledge are also likely to show such bias. Any new selection test (whether based solely on a knowledge base or not) needs to include an assessment of such biases. Emery and Bell (2009) do state that 'work is currently underway to assess the equity of the BMAT' and UKCAT have already published detailed data on the relationship of its tests to a wide range of social variables (UKCAT 2008a, 2008b, 2009), and it seems at least probable that BMAT scores will probably be higher in males, and those from private sector schools or attending institutions with more experience of taking BMAT.

Comment

The analyses in the Emery and Bell (2009) paper are typical of a number of analyses conducted in the medical education...
Validity of the BMAT


References


JOANNE L. EMERY & JOHN F. BELL
Cambridge Assessment, UK

Medical admissions tests have high stakes for both takers and users. Sophisticated criteria should indeed apply to their design and operation. We constantly emphasise the need for validation effort and have argued for this to be extended and enhanced within the UK setting. We are therefore extremely disappointed in the approach taken in McManus et al., which appears to have failed to engage with the detail of work completed to date.

Specifically, their paper:

- ignores the fact that we (Emery & Bell 2009) stated clearly, as our major finding, that Section 2 of the BMAT is the stronger predictor of early medicine course performance for these cohorts. Much of our discussion was devoted to this. McManus et al. present this as a novel conclusion resulting from their work. It is not.
- reiterates other points stated in our paper (e.g. the need for a test to be shown to be unbiased before it can be accepted as valid) and identifies problems that were already acknowledged (e.g. that of comparing the odds ratios for different years due to differences in the score ranges of the test).
- ignores the full data we presented. Means and standard deviations were not given but, to aid in the interpretation of results (particularly the logistic regression plots), histograms were presented displaying the BMAT scores of accepted and rejected applicants in each cohort. These were available in a supplementary table (S1) that we twice referred to in our paper.
- implies that our analyses yielded misleading and inflated results – yet techniques suggested in McManus et al. decrease sensitivity (Table 1) and would result in elevated coefficients (if statistical corrections for reliability were applied). In presenting our correlations uncorrected and separated by cohort, low coefficients were evident in places and the variability typically found in this field (see review by Julian 2005) was not masked by a ‘more reliable’ value. We feel that the average r values presented in Table 1 of McManus et al. actually create a rosier impression, although interpreting them according to traditional effect size definitions is overly-stringent for reasons we stated in our introduction. Their presentation of Q values in order to suggest that an average value would often suffice is unconvincing, given that Q has low power to detect heterogeneity when numbers are small – 4 in these analyses (Huedo-Medina et al. 2000).
- presents contradictory definitions of bias. Their introduction correctly states that bias cannot be defined as differences in mean performance between groups, yet elsewhere their argument relies on this definition. Emery et al. (2010) examines bias in considerable detail in order to establish more advanced analytic practice. DIF analysis, as an internal measure of bias, is insufficient for an admissions test (Camilli 2006).
- misconstrues BMAT Section 2 as ‘intended to be at the level of GCSE science’ or ‘designed to measure science ability at GCSE’. Section 2 requires the application of scientific knowledge encountered up to this stage – it is designed to be challenging for even the most able A-level candidates.
- erroneously suggests the use of partial correlation. The moderate correlation between BMAT Sections 1 and 2 does not mean that they measure the same thing – they do not. Biology, Chemistry and Mathematics A-level grades may correlate highly – would the authors therefore suggest the same statistical treatment of these in assessing their utility for medical selection?
- concludes by demanding an array of validity evidence that is beyond the scope of a single journal article. Many requirements must be met in the process of validation for a selection test but our paper claims no more than to provide

Correspondence: Joanne L. Emery, Research Division, Cambridge Assessment, 1 Regent St., Cambridge CB2 1GG, UK. Tel: 01223 55384; fax: 01223 552700; email: emery.j@cambridgeassessment.org.uk
evidence of predictive validity, for a single institution and course, and should be judged only against this claim. The demonstration of predictive validity must be given priority for tests like the BMAT if they are to be accepted as fit for purpose.

- underestimates the challenge faced by medical schools in developing appropriate selection processes for all candidates and not just those with A-levels.

This last point is extremely important. Prior to the awarding of A* grades at A-level in 2010, almost all candidates presented with three A grades, or were overseas candidates presenting a range of alternative qualifications. There are limitations, therefore, in using qualifications-related information for admissions. We argue that the ceiling in A-level grades in our datasets indicates that incremental validity was obtained by the BMAT. The authors’ suggestion of using the number of A-level grades as a source of variance is questionable – this will clearly be related to an applicant’s school type. Using the A* grade and/or unit marks may hold potential, but only in relation to UK candidates.

An extended response to McManus et al. is available on the Cambridge Assessment website: http://www.cambridgeassessment.org.uk/ca/Our_Services/Research/Conference_Papers

Declaration of interest: Both authors work for Cambridge Assessment, who own and administer the BMAT.

References


Response to comments by Emery and Bell, Medical Teacher 33(1): (this issue)

I. C. MCMANUS1, EAMONN FERGUSON2, RICHARD WAKEFORD2, DAVID POWIS3 & DAVID JAMES5

1University College London, UK, 2University of Nottingham, UK, 3University of Cambridge, UK, 4University of Newcastle, Australia, 5Queens Medical Centre, UK

We thank Emery and Bell (E&B) for their detailed comments which, to our eyes at least, reinforce most of the criticisms that were raised in our paper. E&B may say that they were “extremely disappointed in the approach taken” in our analysis, but our paper was written in large part because on reading Emery and Bell (2009) we could not extract the information that would convince us on psychometric grounds that BMAT was an effective test. E&B of course hold all the cards, having the original data, and we, like any outsiders, can only carry out secondary analysis of what they choose to provide. We believe we have interpreted carefully the numbers that E&B presented, we have weighed them in the balance as best we can, and to a large extent have found them wanting. BMAT is a high-stakes assessment run by “one of the world’s largest assessment agencies” (www.admissiontests.cambridgeassessment.org.uk/adt/bmat) and if we, as educationalists, cannot convince ourselves of the acceptability of the test, we find it difficult to know the objective, scientific grounds on which candidates, parents, teachers, and medical schools should also be willing to accept it.

Before responding to specific matters raised by E&B, we will firstly discuss the one key issue which they chose to ignore in their response.

Why have we still not been told the reliability of BMAT?

E&B raise a number of detailed comments on our paper, but perhaps most striking about the E&B response is what is absent from it. In particular, despite our paper mentioning on multiple occasions the need for measures of reliability, and the fact that Emery and Bell (2009) did not provide them, the reliability coefficients for BMAT Section 1 and 2 are still not presented by E&B. Reliability is largely a function of the square root of the number items in a test (n), so that in a fixed population (such as medical school applicants) it is n that relates to reliability. BMAT Sections 1 and 2, have only 35 and 27 items. A comparable test for a comparable population, the total score from UKCAT, has 175 items and a reliability of 0.86 (UKCAT, n.d.). The Spearman–Brown formula, predicts that tests with 35 and 27 items (such as BMAT Sections 1 and 2), would have reliabilities of .55 and .48. Such values are unacceptably low for high stakes tests, particularly as it is only Section 2, the test with the lower reliability that is reported to have any predictive validity. Of course, it may be that BMAT has peculiarly well-discriminating items, so that the reliabilities are far higher than our crude estimates. Accordingly, we would like to see the reliability of the BMAT published.

Validity

Two of our major concerns are construct validity and predictive validity of a test. We state clearly that without construct validity – identifying exactly what a test is measuring – there is a danger of reinventing the wheel. If a selection measure is an index of scientific knowledge, that is measuring the same construct as GCSE or A level science, then BMAT is but an additional test of the same construct. Construct validity is therefore crucial. E&B question whether section 2 is an index of GCSE or A level science. The only way to resolve that issue is to provide data on the association between GCSE and A level science and Section 2. It may indeed be the case that all candidates have straight A grades at A level (but that seems unlikely given that many applicants apply before A-level, and not all predictions of straight As manifest as such). Even if it were the case, comparisons could be made to GCSE scores, or better still, a separate validity study conducted of BMAT in a larger cohort of A level students with a wider range of A level scores. It is of interest that UKCAT does correlate with A-level grades (James et al. 2010). Of the types of validity required to establish a test, construct validity should be one of the first to be established. A related issue is of establishing incremental predictive validity, since it is necessary to show that the new test predicts above and beyond existing and established selection methods. Even if A levels are at ceiling, then GCSE results offer at least one possible alternative. What we are suggesting is the particular need to explore and include many predictors in a single multivariate model.

Meta analysis of correlations with examination performance

We do not imply that the analyses of Emery and Bell (2009) lead to inflated results. In fact, our meta-analysis shows that...
their correlations, particularly given our Q values, are generally consistent. We do suggest that corrections for unreliability should also be applied. This is not to create a ‘rosier’ picture, but to provide the reader with full information about the potential limits to validity, and the extent to which the current correlations are low because of low reliability (due to tests which are fairly short, a problem that is correctable) or to a low intrinsic correlation (which would require a very different approach). Corrected scores should, as E&B state, be presented with the appropriate caveats. In this way the precision of the point estimates can be judged. Indeed, this is why in a number of places we emphasize the need for confidence intervals.

**Partial correlations and shared variance in predictors**

We used partial correlations to control for the shared variance between Sections 1 and 2. This is not to suggest that the two sections are measuring the same thing, but that they share variance. This shared variance may be due to error, similar response styles etc., in just the same way that the shared variance between different A-level subjects may be due to good teaching across all subjects by a particular school, or a common intellectual aptitude, motivation or study methods on the part of a student which result in similar grades in all the subjects. The point is to establish the independent or joint effect of each test component and to do this we need to control the shared variance (it is always informative to partition shared variance among predictor variables to develop a multivariate model). While Emery and Bell (2009) do state that Section 2 has stronger correlations with the outcomes than Section 1, they provide no statistical evidence for this. What we aimed to do was to explore the claim statistically, and we showed not just that Section 2 had stronger associations with outcomes, but that the associations with Section 1 were reduced to being either non-significant or to being very small. The statistical exploration has resulted in far clearer conclusions.

**Effect sizes**

We state that the effect sizes, by conventional criteria, are small to medium. With respect to interpreting the practical significance of effect size, the guidelines of Prentice and Miller (1992) are informative. Small effect sizes should be considered impressive, a) when the intervention is minimal, or b) when the outcome is difficult to influence. Hence within medicine a minimal intervention (e.g., aspirin) with a small but significant effect in reducing a difficult to influence outcome (e.g., risk of future cardiovascular events) potentially has important public health implications (Steering 1988). Whether BMAT is a minimal intervention is not quite so clear, and neither is it clear that the outcome measure of future exam performance is something that is difficult to influence. In this case we consider that this is a small effect size and should be described as such. It should also be noted that one type of exam performance (the selection test) is being used to predict performance on another type of exam (end of year exams). Issues of common method variance are therefore potentially problematic, and we do wonder what the results would be if modelled using a structural model with correlated errors in the predictor and outcome (and simple correlations and logistic regression assume the predictor measures are error free).

**Conclusions**

The aim of our original commentary was to highlight the psychometric criteria (e.g., construct validity, incremental predictive validity, and reliability) and analyses that are needed to establish a selection test as one providing additional data that are useful and informative in the selection context. We feel that our original paper still makes those points, and we hope that our reply to E&B serves to strengthen our original comments, and clarifies the issues for readers of both papers. Our call is for a fuller presentation of data and more robust statistical analyses of all studies examining the predictive validity of medical selection tests. Indeed, we used the Emery & Bell paper to make a number of key observations about what constitutes good scientific evidence to support the use of aptitude tests in medical selection. For example, were E&B to put their raw data in the public domain then that suggestion would of course be more easily satisfied.

BMAT is a high stakes tests for which UK candidates currently pay £42.50 to take. Surprisingly, the numbers of candidates actually taking the examination appear not to be published; the brief reports for 2004 to 2009 describing results only in percentage terms (www.admissionstests.cambridgeassessment.org.uk/adt/bmat/about#Results). That surely is less than good practice. Numbers were published for 2003, when about 5000 candidates took the exam, and it is unlikely that fewer have taken it since. For an estimated income of about £200,000 per annum, there should be an obligation upon Cambridge Assessment itself, rather than E&B alone, to provide proper analyses of BMAT.

**Declaration of interest:** ICM, DP and DJ have been involved with UKCAT (United Kingdom Clinical Aptitude Test).

**References**


