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

I. C. McManus1, EAMONN FERGUSON2, RICHARD WAKEFORD3, DAVID POWIS4 & DAVID JAMES5
1University College London, UK, 2University of Nottingham, UK, 3University of Cambridge, UK, 4University of Newcastle, Australia, 5Queens Medical Centre, UK

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 (Carraquillo & 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.
Skill' section and it assesses problem solving, argument

The BMAT is in two sections. Section 1 is the 'Aptitude and
Skill' section and it assesses problem solving, argument comprehensiveness, 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).

The BMAT

The BMAT is in two sections. Section 1 is the 'Aptitude and Skill' section and it assesses problem solving, argument

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>Meta-analyses for individual course components</th>
<th>Meta-analyses for total marks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Functions of the body</td>
<td></td>
</tr>
<tr>
<td>Section 1</td>
<td>Year 1</td>
<td>Molecules in medical science</td>
<td>Total</td>
</tr>
<tr>
<td>Homeostasis</td>
<td></td>
<td></td>
<td>$r = 0.17^*; Q = 2.06$ ns</td>
</tr>
<tr>
<td>Biology of disease</td>
<td>$r = 0.16^*; Q = 0.88$ ns</td>
<td>$r = 0.14^*; Q = 0.99$ ns</td>
<td>$r = 0.20^*; Q = 0.99$ ns</td>
</tr>
<tr>
<td>Year 2</td>
<td>Human reproduction</td>
<td>Neurobiology and human behaviour</td>
<td>Total</td>
</tr>
<tr>
<td>Biology of disease</td>
<td>$r = 0.33^*; Q = 5.80$ ns</td>
<td>Neurobiology and human behaviour</td>
<td>Total</td>
</tr>
<tr>
<td>Year 2</td>
<td>Human reproduction</td>
<td>Neurobiology and human behaviour</td>
<td>Total</td>
</tr>
<tr>
<td>Science</td>
<td></td>
<td>$r = 0.19^*; Q = 0.99$ ns</td>
<td>$r = 0.20^*; Q = 0.99$ ns</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 ($Q(999) = 5.3, p < 0.0001$; 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 ($Q(999) = 2.37, p = 0.017$). 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

Differential 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 A, 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 . . . , [T]he 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
literature which do not meet basic psychometric standards. This is important as cognitive tests of this type are being used to make important decisions about who will be the doctors of the future. The use of zero-order analyses and correlations in the absence of reliability data, and failure to take into account the inter-relatedness of measures using structural modelling, hierarchical regression or multi-level modelling, inevitably mean that a distorted picture may be presented. Even the simple partial correlations presented here show that almost all of the predictive power of BMAT resides in Section 2, which assesses Scientific Knowledge and Applications, a section which resembles a standard academic test, and therefore potentially open to similar problems of bias. We suggest that all future papers on the validity of selection tests include (1) means and SDs for selected and non-selected candidates; (2) data on reliability (internal and test-retest); (3) construct validity; (4) predictive incremental validity against traditional selection tools and (5) estimates of bias due to sex, ethnicity and SES. Our aim in this article has been to show that proper reporting of evidence is necessary, and that the selective reporting of predictive validity data can hide a range of crucial issues, making a test appear more successful and less problematic than it is. Candidates for admission to medical school, whose future careers depend on such tests, deserve better, as perhaps also do their future patients.

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

Contributions: The first draft of the manuscript was written by ICM, and a much modified second draft was authored by EF. All authors subsequently discussed the manuscript and contributed to the final draft.

References


