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

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

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**


