Extended Response to I C McManus, Eamonn Ferguson, Richard Wakeford, David Powis and David James (2011). Predictive validity of the BioMedical Admissions Test (BMAT): An Evaluation and Case Study. Med Teach 33 (1): 53-57.

John F Bell and Joanne L Emery

January 2011

Introduction

To fully understand the background to the McManus et al. (2011a) critique of our work on the predictive validity of the BMAT (Emery and Bell, 2009) we invite readers to follow the link to their previous paper (McManus et al., 2005) *and* the response by Bell (2005). This puts the McManus et al. (2011a) paper into its proper context.

In their critique of our work, the authors claim to be advocating good practice and, similarly motivated, we now wish to put forward our own understanding of some of the relevant principles of psychometric theory. We wish to further investigate the difficulties associated with analysing data affected by selection and of understanding test bias. The following points are in addition to our short response in *Medical Teacher* (Emery and Bell, 2011).

Incorrect Calculation

McManus et al. (2011b) state that: "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." In their paper, the authors present results based on their re-analyses of our correlation coefficients (see Table 1 of McManus et al. 2011a). Unfortunately, they do not give exact details of their calculations but it is difficult to see how they arrived at the overall values given for correlations between BMAT Section 2 and Neurobiology and Human Behaviour and between BMAT Section 2 and Functional Architecture of the Body. In the former case, they produce some form of weighted average with the value 0.25 by combining four correlations of 0.35, 0.24, 0.24 and 0.27. In the latter case, they obtained a value of 0.42 from correlations of 0.40, 0.26, 0.41, and 0.16. All the other overall correlations in their table appear similar to the weighted means of the values we gave in Emery and Bell (2009). Given the approximate equality of numbers across our four cohorts, it seems certain that some of the calculations in Table 1 are erroneous.

Reliability

It is our policy to focus our empirical effort on the predictive validity of a test once outcomes data is available to make this possible. Reliability information for the BMAT has long been available on the Cambridge Assessment website (Wilmott, 2005) and appears readily in a Google search. It is common practice (AERA, APA, NCME Standards for Educational Research, 1999: p31) for test developers to report basic test statistics and these will be reported for the BMAT as a matter of course following each exam session. These will be available for the November 2010 BMAT session on the Cambridge Assessment Admissions Testing website in February 2011. The most usual measure of reliability is Cronbach's alpha, which measures internal consistency. BMAT sections 1 and 2 each have typical alpha values of .65. Although higher than the values of .55 and .48 estimated by McManus et al. (2011b),

Nunnally's (1978) guidelines suggest that these are too low if a section were to be used on its own in the selection process. Schmitt (1996), however, considers this guideline to be shortsighted. One of the reasons he gives is:

"Classic reliability theory also holds that the upper limit of validity (the relationship between a predictor and criterion) is the square root of the reliability of the criterion or outcome variables rather than 1.00, which is the upper limit of a Pearson correlation. The concern then is that the true correlations involving a predictor and an unreliable outcome variable will be seriously attenuated (i.e. underestimated) because of inadequate criterion reliability rather than any lack of real or true relationship. In considering the implications of these findings for expected validity, it can be seen that with reliability equal to .70, validity has an upper limit of .84 (i.e., the square root of .70) as opposed to 1.00. Even with reliability as low as .49, the upper limit of validity is .70. When a measure has other desirable properties, such as meaningful content coverage of some domain and reasonable unidimensionality, this low reliability may not be a major impediment to its use."

We argue that maintaining content coverage is more important for the validity of the BMAT than is maximising its internal consistency. BMAT Section 2, for example, contains items on biology, chemistry, physics and mathematics. Its reliability would undoubtedly be higher if it contained only mathematics items but its validity for medical student selection would be seriously impaired. Cronbach and Shavelson (2004) note that:

"The alpha formula is not strictly appropriate for many tests constructed according to a plan that allocates some fraction of the items to particular topics or processes. Thus, in a test of mathematical reasoning, it may be decided to construct 20% of the items around geometric shapes. The [correlations among items of] several forms of the test that could be constructed by randomly sampling geometric items will be higher than the correlation among items in general."

Although it is true that if a test is completely unreliable it can have no predictive validity, it does not automatically follow that increasing reliability will increase validity. Feldt (1997) demonstrated how it is possible for validity to decline as reliability increases. From a pragmatic perspective, it is more important that a test predicts subsequent performance than reaches some arbitrary level of internal consistency but does not predict.

Increasing the number of items in each of the BMAT sections would raise their alpha values but this must be weighed against considerations such as the time needed to complete the test and the cost to candidates. McManus et al. (2011a) state "it is also clear, that to a first approximation, reliability is proportional to the square roots of the number of items." This relationship occurred by chance in the UKCAT data presented. It is impossible to infer the reliability of a test solely on the basis of its test length. Cronbach's alpha can be formulated as being the true score variance divided by the total variance (the sum of the true score variance and the error variance). Whilst the error variance is inversely related to the number of items, the true score variance is not related to the number of items at all. This means that alpha can vary considerably depending on the distribution of true scores in the study so it cannot be related straightforwardly to the number of items.

The Effects of Selection

When evaluating a test that is already in use as a piece of admissions evidence, the statistical effects of selection need to be taken into account. Selection has two

consequences. Firstly, the correlations between test scores and future performance are affected by range restriction because there is no data for the weakest (rejected) candidates. Secondly, both regressions and correlations are affected by the fact that any candidates accepted with low test scores are likely to be atypical of low-scorers in general. It is reasonable to assume that candidates accepted with low BMAT scores are likely to perform better than would those rejected with the same scores, because they were admitted on the basis of other admissions evidence that compensated for their low BMAT scores. This leads to a tendency for regression slopes to be too shallow and correlations too low. Although there has been research into correcting for these problems (Sackett and Yang, 2000), we took the most conservative approach and reported uncorrected correlation coefficients in Emery and Bell (2009). The fact that the BMAT is used in a compensatory selection process means that the simplest corrections for range restriction were not applicable to the data we used, as we explained in our paper. Linn (1982) noted that:

"Correlations are not routinely corrected for range restriction in predictive validity studies— in large part, because the formula depends on assumptions of homoscedasticity and linearity that are violated to an unknown degree. Therefore, the conservative position is not to correct for range restriction. Such conservatism is laudable, but it is generally desirable to obtain and report the corrected estimate along with the uncorrected correlation."

However, the problem with presenting a corrected correlation is that using an incorrect method would lead to justified criticism and as the selection process increases in complexity the outcome would require too many challengeable assumptions. We therefore presented uncorrected correlations, along with the caveat that these were a worst case scenario, rather than risk using a correction that was not strictly appropriate.

A further problem with the complexity of the selection process is that the reduction in the correlation for a given source of admissions evidence is proportional to the weight it is given in the selection process (Linn and Dunbar, 1982; Bell, 2007). This creates particular problems with partial correlations calculated from zero-order correlations. Given that the correlations between the BMAT sections and the outcome variable are underestimates and that the degree of underestimation varies, using them in the calculation of partial correlations is not to be recommended. This is not the only problem with advocating the use of the partial correlations. McManus et al. (2011a and b) overlooks the literature (Burks, 1926a and b; Stouffer, 1936; Wolins, 1967, Gordon, 1968; Lord,1974) that demonstrates the inadvisability of calculating partial correlations for variables subject to measurement error.

Most importantly, as Linn and Werts (1973) point out:

"Ignoring measurement errors is much more serious when dealing with partial correlations than when dealing with simple zero-order correlations. In the latter case, we know that the effect of errors of measurement ….. is to reduce the absolute value of the zero-order correlation between the fallible measures. As Lord (1963) has pointed out, however, we cannot ordinarily know the effect of such errors of measurement on a partial correlation. Errors of measurement can increase or decrease the magnitude of a partial correlation and may even result in a partial correlation of a different sign."

McManus et al. (2011a) also apply Cohen's criterion for assessing the size of the correlation coefficient. It is interesting to note what Cohen (1988) actually wrote:

"The preceding serves as an introduction to the operational definitions of 'small', 'medium', and 'large' ES [effect size] as expressed in terms of r [correlation], offered as a convention. A reader who finds that what is here defined as "large" is too small (or too large) to meet what his area of behavioural science would consider appropriate standards is urged to make more suitable operational definitions."

Researchers in the area of admissions tests have made such operational definitions, which were quoted, along with justification, in our paper. The values are lower than those of Cohen, firstly, for the reasons described above and secondly, for the fact that even small correlations can make a considerable difference to the quality of candidates selected. This is because of the concept of predictive efficiency. In 1939, Taylor and Russell noted that the predictive efficiency of a validity coefficient will be a function of the proportion of individuals considered satisfactory on the basis of some criterion measure and the proportion of the tested group which is selected. Therefore the forecasting value may be considerably higher than is indicated by the correlation coefficient. Using what would, for most selection processes, be overly-simplistic assumptions, Taylor and Russell produced tables that show the effect of these factors.

Consider an example where the percentage of applicants who would have been considered satisfactory had they been selected was 70% and the selection ratio (the number of places divided by the number of applicants) was 0.2. If a selection test with a correlation of 0.15 was used as the sole selection criterion then 77% of the admitted applicants would be deemed satisfactory. Squaring the correlation would suggest that the test explains only 2.25% of the variation. It is also worth noting that these tables are based on the correct value of the correlation and not the underestimate resulting from range restriction effects. Obviously, Taylor and Russell's tables are based on an over-simplification. However, Bell (2007) extended their ideas with a simulation and found large gains in predictive efficiency when two criteria were used in a selection process.

The Nature of Test Bias

McManus et al. (2011a) cite a number of their papers that purport to show biases. These actually only show group differences in test performance. The authors do acknowledge at one point that group differences in test performance may reflect real group differences rather than bias. However, the popular misconception that group differences indicate a bias in the measurement tool is reinforced throughout their paper in statements such as:

"…based on the claim that traditional 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…"

and:

"Given that standard academic assessments are, for a host of reasons, influenced by sex, ethnicity, socioeconomic background and so on (Powis et al. 2007; McManus et al. 2008), predictive tests that are based on knowledge are also likely to show such bias."

Such statements can easily lead to misunderstandings about the fairness of admissions tests should group differences in performance be found. We believe that the Standards for Educational and Psychological Testing (AERA, PA, & NCME, 1999) offer several valuable insights here. The guidelines state that:

“the idea that fairness requires equality in overall passing rates for different groups has been almost entirely repudiated in the professional testing literature” (p. 74).

It is further stated that:

“most testing professionals would probably agree that, while group differences in testing outcomes should in many cases trigger heightened scrutiny for possible sources of test bias, outcome differences across groups do not in themselves indicate that a testing application is biased or unfair” (p. 75).

The measurement of test bias is discussed in more detail in Emery, Bell and Vidal Rodeiro (2011), which examines the predictive equity of the BMAT rather than simply investigating group differences in performance. Obviously it takes time and space to deal with this issue thoroughly and not all aspects of validity can be covered in a single journal paper.

Differential item function (DIF) analysis is not sufficient to identify problems with bias. This only detects potentially biased items. If all the items in a test were biased, DIF analysis would not detect the bias (as all the items would behave equally badly). Even when some of the items function differentially this is still not necessarily evidence of bias. For example, in Science GCSEs female candidates tend to perform less well on physics items but a test without them would not be valid. DIF analysis should always involve a second step that requires judgement about whether the difference results from a valid or an invalid source of difficulty.

Inaccurate Portrayal of the BMAT and its Function

As noted in the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999), a test is appropriately evaluated in terms of how well it meets the claims made for it. It is not reasonable to fault the test for (a) not measuring characteristics other than those it purports to measure or (b) not being the sole and complete determinant of job or educational success.

McManus et al. (2011a) state that:

“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.”

This is not the case. The BMAT’s stated aims are:

“BMAT was developed by Cambridge Assessment in response to a request by academics from some of the top medical and veterinary schools in the UK for an assessment that would:

Enable them to differentiate between applicants who appear to be equally well qualified and suited to the course.
Provide a way of assessing the potential of students who have a range of qualifications.” (http://www.bmat.org.uk).

The word ‘potential’ was intended to mean no more than the OED dictionary definition:

“having or showing the capacity to develop into something in the future”

In the case of the BMAT this is the capacity to succeed on a medical degree. Section 1 of the BMAT is indeed called “Aptitude and Skills”. However, McManus et al. (2011a) read too much into this. The OED gives the definition of an aptitude test as:

“ a test designed to determine a person's capacity in any given skill or field of knowledge.”

The OED definition of the word ‘aptitude’ is:

“The quality of being fit for a purpose or position, or suited to general requirements; fitness, suitableness, appropriateness.”

It can also be taken to mean “Natural capacity to learn or understand; intelligence, quick-wittedness, readiness.”

However, it is clear from the more detailed BMAT Section 1 specifications that the latter definition is not what was intended:

“This section tests generic skills in problem solving, understanding argument and data analysis and inference.”

The nature of BMAT Section 2 (Scientific Knowledge and Applications) is also inaccurately portrayed in McManus et al. (2011a):

“…section 2, which the developers say is designed to measure science ability at National Curriculum Key Stage 4 (i.e. GCSE)...”

It is described on the BMAT website as follows:

“This section tests a candidate’s ability to apply scientific knowledge encountered in non-specialist school science and maths courses, up to and including National Curriculum Key Stage 4. The BMAT is designed to supply evidence that will help to identify candidates who are ready to start very intense medical courses.”

The difference between BMAT Section 2 and a higher tier Additional Science GCSE can quickly be seen by inspecting a past BMAT paper and a GCSE paper. It is very clear that the BMAT section requires much higher levels of problem solving skills than the GCSE. The aim is to restrict the science knowledge so that it is not a hurdle to the assessment of advanced reasoning.

McManus et al. (2011a) also state that the BMAT has two sections when it has three. This is clearly stated in our paper, on the BMAT website and in all materials relating to the test. Marks for the third section were not included in Emery and Bell (2009), as we explained, because they are not used in the selection process at the University of Cambridge. Copies of the essays are sent to the admissions tutors who may use them in the interview. The statement that both sections are assessed using MCQs (multiple choice questions) is also inaccurate. They are composed mostly of MCQs but can require numerical answers, as was also stated in our paper.

McManus et al. (2011a) apply the theory of fluid and crystallized intelligence to the BMAT, stating that “Section 2 is explicitly an index of crystallised intelligence but, from the description, so in part is section 1” and that “Section 1… is largely a test likely to reflect fluid intelligence,…” They also state that “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')."

The BMAT and its predecessor, the MVAT, were not developed as intelligence tests but rather with the pragmatic aim of identifying students who would be suitable for the intense scientific study involved early in these medical courses. The students accepted onto the courses that use the BMAT are expected to be ready to make rapid progress on them. Both sections 1 and 2 involve, to differing degrees, a combination of knowledge recall and more fluid intelligence (e.g. problem solving / reasoning skills).

Conclusion

We have welcomed the chance to address the issues raised in McManus et al. (2011a; 2011b). The purpose of our first paper on the BMAT (Emery and Bell, 2009) was to demonstrate the predictive validity of the test in a transparent manner. We believe that this aspect of validation work has to be given priority for admissions tests, particularly in response to continued calls in the medical literature for such evidence. We agree that more is required to establish any selection test as psychometrically sound, as was indeed stated in our paper. The BMAT undergoes a continuous process of evaluation in order to fulfil our aim of providing an educationally useful service to higher education and our early work should be regarded as part of a wider dissemination strategy.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999) *Standards for educational and psychological testing*. Washington, DC: National Council on Measurement in Education.

Bell, J.F. (2005) *The case against the BMAT: Not Withering but Withered?* Available online at http://www.cambridgeassessment.org.uk/ca/digitalAssets/113949_The_case_against_the_BMAT.pdf

Bell, J.F. (2007) Difficulties in evaluating the predictive validity of selection tests. *Research Matters: A Cambridge Assessment Publication,* 3, 5-10. Available online at http://www.cambridgeassessment.org.uk/ca/digitalAssets/169423_Research_Matters_3_Jan_2007.pdf

Burks, B.S. (1926a) On the inadequacy of the partial and multiple correlation technique: Part I. *Journal of Educational Psychology,* 17, 532-540.

Burks, B.S (1926b) On the inadequacy of the partial and multiple correlation technique: Part II. *Journal of Educational Psychology*, 17, 625-63.

Cronbach, L.J. and Shavelson, R.J. (2004) My Current Thoughts on Coefficient Alpha and Successor Procedures. *Educational and Psychological Measurement,* 64, 391-418.

Cohen, F.C. (1988) *Statistical Power analysis for the behavioural Sciences*. Second Edition. Hillsday, HJ: Lawrence Erlbaum Associates.

Emery, J.L. and Bell, J.F. (2009) The predictive validity of the BioMedical Admissions Test for pre-clinical examination performance. *Medical Education*, 43, 557-564.

Emery, J.L. and Bell, J.F. (2011) Comment on I.C. McManus, Eamonn Ferguson, Richard Wakeford, David Powis and David James (2011). Predictive validity of the BioMedical Admissions Test (BMAT): an evaluation and case study. *Medical Teacher,* 33, 58-9.

Emery, J.L., Bell, J.F. and Vidal Rodeiro, C.L. (2011) The BioMedical Admissions Test for medical student selection: Issues of fairness and bias. *Medical Teacher,* 33, 62-71.

Feldt, L.S. (1997) Can validity rise when reliability declines? *Applied Measurement in Education*, 10, 377-387.

Gordon, R.A. (1968) Issues in multiple regression. *The American Journal of Sociology*, 73, 592-616.

Linn, R.L. (1982) Admissions testing on trial. *American Psychologist,* 3, 279-291.

Linn, R.L. and Dunbar, S.B. (1982) Predictive validity of admissions measures: Correction for selection on several variables. *Journal of College Student Personnel*, 23, 222–226.

Linn, R.L. and Werts, C.E. (1973) Errors of Inference Due to Errors of Measurement. *Educational and Psychological Measurement*, 33, 531-543.

Lord, F.M. (1963) Biserial estimates of correlation. *Psychometrika,* 28, 81-85.

Lord, F.M. (1974) Significance test for a partial correlation corrected for attenuation. *Educational and Psychological Measurement*, 34, 211-220.

McManus, I.C., Powis, D.A., Wakeford, R., Ferguson, E., James, D. and Richards, P. (2005) Intellectual aptitude tests and A levels for selecting UK school leaver entrants for medical school. *British Medical Journal*, 331, 555-559.

McManus, I.C., Elder, A.T., De Champlain, A., Dacre, J.E., Mollon, J. and Chis, L. (2008) Graduates of different UK medical schools show substantial differences in performance on MRCP(UK) Part 1, Part 2 and PACES examinations. *BMC Medicine*, 6, 5.

McManus, I.C., Ferguson, E., Wakeford, R., Powis, D. and James, D. (2011a) Predictive validity of the BioMedical Admissions Test (BMAT): An evaluation and case study. *Medical Teacher*, 33, 53-57.

McManus, I.C., Ferguson, E., Wakeford, R., Powis, D. and James, D. (2011b) Response to Comments by Emery and Bell, Medical Teacher (33(1): (this issue). *Medical Teacher,* 33, 60-61.

Nunnally, J.C. (1978) Psychometric theory (2nd ed.). New York: McGraw-Hill.

Powis, D., Hamilton, J. and McManus, I.C. (2007) Widening access by changing the criteria for selecting medical students. *Teaching and Teacher Education*, 23, 1235–1245.

Sackett, P. R. and Yang, H. (2000) Correction for range restriction: An expanded typology. *Journal of Applied Psychology*, 85, 112–118.

Schmitt, N. (1996) Uses and abuses of coefficient alpha. *Psychological Assessment,* 8, 350-353.

Stouffer, S.A. (1936) Evaluating the effects of inadequately measured variables in partial correlation analysis. *Journal of the American Statistical Association*, 31, 348-360.

Taylor, H.C. and Russell, J.T. (1939) The relationship of validity coefficients to the practical validity of tests in selection: Discussion and tables. *Journal of Applied Psychology*, 23, 565–578.

Willmott, A. (2005) *Thinking Skills and Admissions. A report on the Validity and Reliability of the TSA and MVAT/BMAT Assessments.* Available online at http://www.cambridgeassessment.org.uk/ca/digitalAssets/113977_Thinking_Skills___Admissions_a_report_on_validity.pdf

Wolins, L. (1967) The use of multiple regression procedures when the predictor variables are psychological tests. *Educational and Psychological Measurement,* 27, 821-827.