Author's response to reviews

Title: Even one star at Alevel could be "too little too late" for medical student selection.

Authors:

Chris McManus (i.mcmanus@ucl.ac.uk)
Kath Woolf (k.woolf@medsch.ucl.ac.uk)
Jane Dacre (j.dacre@medsch.ucl.ac.uk)

Version: 2 Date: 14 January 2008

Author's response to reviews:

Responses to reviews. We are grateful to the reviewers for their thoughtful comments, and we hope we have responded adequately to them. In the following responses, for clarity, the original comments are in Roman script, and our responses are in italics.

Reviewer's report
Title: Even one star at Alevel could be "too little too late" for medical student selection.
Version: 1 Date: 16 September 2007
Reviewer: Sandra Nicholson
Reviewer's report:

General

I enjoyed reading the authors' analysis of the ability of current A Level grades to aid medical student selection with their cautionary prediction of how introducing A* grades will only facilitate discrimination by academic merit for approximately the next decade. Whilst previous work of the first author has demonstrated the predictive validity of success at A Level with medical school and postgraduate examinations I believe the authors were right to attempt to demonstrate again this assertion in the light of recent A Level inflation. The data concerning year 3 assessments (MCQ and OSCE) appear to corroborate with these previous findings. However we do not know, and the authors are assuming, that students with A*/A** will perform better in medical school examinations.

Response: This is a fair point, and reminds all educationalists that validity is not something that can be assumed for all tests for all times. Tests and times change, and validity should be routinely monitored for all selection methods.
Having said that, until recently there was a surprising dearth of evidence that A levels were actually valid predictors of medical school outcome. We have added a brief comment on this in the discussion.

The authors are right to call for further research into the relationship between academic success as routinely measured by A Level grades and performance at medical school, but I would also add more importantly as practising doctors.

Response: We agree entirely with this and have added a line to the discussion saying, "Predictive validity should also be examined broadly, not merely for examination success, but for other aspects of real world clinical practice in working doctors."

This introduces my main philosophical point that whilst introducing A* and perhaps A** grades facilitates discrimination between applicants for medical courses, well demonstrated by this paper, in terms of who to reject it does not necessarily help select candidates on the basis of who will go on to be the best and happiest doctors. Surely there must be a ceiling on the relationship between academic success and performance at medical school and certainly with medical practice.

Response: Whether or not there is a ceiling to the predictive ability of A-levels or other educational measures is an empirical question and an important one. In unpublished data we have never found any evidence of non-linearity in the predictive validity of A-levels. Perhaps with the very large databases being collected by consortia such as UKCAT then the question will be properly unanswerable.

The paper does not acknowledge that most medical schools take into consideration, some highly, other measures of suitability for medical training. To be fair A Levels remain the most heavily weighted selection tool currently but with their discussed limitations it is time to seek other reliable and valid methods of selecting tomorrow's doctors. Therefore whilst I agree with the authors A* may be "too little too late" I would also add "too narrow" and urge the medical selection community to consider alternative selection tools alongside the traditional measurements of academic success.

Response: We accept this argument, and acknowledge fully that some medical schools do take a range of other measures into account. Having said that, good evidence of their validity is rare to the point of being almost entirely lacking. At present we think it is reasonable to assert that A-levels are the only selection measure in use in the UK that has proper evidence of long-term predictive validity. We will look forward to similar evidence being provided for other measures.

-------------------------------------------------------------------------------

Major Compulsory Revisions (that the author must respond to before a decision
on publication can be reached)
None

---

Minor Essential Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct)

On page 2

Results

The phrase "A* grades had been already introduced..." makes the meaning of the sentence unclear do they mean "if or had A* grades had been already introduced" ??

Response: We agree this is ambiguous and have changed the wording to, "If it were the case that A* grades had already been introduced, then...".

On page 3

Background: Bearing in mind my general comments above I would ask the authors to consider their use of "unvalidated tests" do they mean tests with no long term predictive validity which is fair comment if contextualised. It is not appropriate to simply say the newly introduced tests for medical student selection are unvalidated when there is evidence of construct validity and reliability amongst internal items and with similar tests in other fields.

Response: We accept that there is evidence of reliability and construct validity for the tests, and since submission of our original manuscript, BMAT has published (non-peer reviewed) data on its website suggesting some predictive validity for its test. We have therefore altered the paragraph to read:

"The result is that many medical schools are now using tests of general intellectual ability for selection for which until recently there was little evidence of predictive validity (although the assessments are reliable and have some construct validity) [1]. Since submission of the first draft of this paper, BMAT has provided some evidence for predictive validity for BMAT=s Section 2 (Scientific Knowledge and Applications) on academic examinations (see www.bmat.org.uk/background.html)."

---

Discretionary Revisions (which the author can choose to ignore)

On page 5

I would question whether the introduction of A*’s was improving the process of selection. Making a process easier does not necessarily make it better and as indicated earlier this paper can, and perhaps should, indicate some of the
competing complexities in student selection.

Response: Easier and better are not indeed equivalent, and we hope we have not implied that they are. However, we do believe that extending the range of an already valid measuring instrument to cover a wider range of candidates can only be an improvement.

Anecdotally some lesser able candidates sitting 4 A Levels may bring their average down. I wonder if the authors would like to comment on the reliability of their data for ALEVUCAS bearing this in mind?

Response: This is a fair point, which is why we only used the three best grades obtained by candidates sitting four A-level grades. There is a risk that candidates have diluted their effort by sitting four A levels when they would have done better by concentrating on three, but in practice there seems to be little evidence of this, presumably because teachers advise such students appropriately. We have never been happy with UCAS' methods of analysing A-levels, but this is probably not an appropriate place to go further in that discussion.

What next?: Accept after minor essential revisions
Level of interest: An article of importance in its field
Quality of written English: Acceptable
Statistical review: Yes, but I do not feel adequately qualified to assess the statistics.
Declaration of competing interests:
I am a director of UKCAT and chair its test development working group. The UKCAT Consortium aim to develop a valid and reliable selection test for medical and dental degree programmes in the UK. These responsibilities do not involve any personal financial gain.

Reviewer’s report
Title: Even one star at Alevel could be "too little too late" for medical student selection.
Version: 1 Date: 21 September 2007
Reviewer: Trudie T Roberts
Reviewer's report:
This paper is interesting and well written as one would expect of these authors. They make the assumption of constant levels of grade inflation. The work by Robert Coe et al suggests that this might be an over estimate. Never the less this paper is well argued in support of greater grade discrimination at the top end of
the distribution and as far as I can tell the models used are appropriate. Although, it should be remembered that I am not a statistician.

Response: In August 2007 the Times reported that Robert Coe had suggested that grade inflation might be different in different subjects, and might be particularly high in maths (a subjects taken by many medical school applicants). As yet we have been unable to find a copy of this report, but are trying.

What next?: Accept without revision
Level of interest: An article of importance in its field
Quality of written English: Acceptable
Statistical review: Yes, but I do not feel adequately qualified to assess the statistics.
Declaration of competing interests:
I declare that I have no competing interests

Reviewer's report
Title: Even one star at Alevel could be "too little too late" for medical student selection.
Version: 1 Date: 16 November 2007
Reviewer: Jeremy Smith
Reviewer's report:

General
This was a well written and informative article, which raises questions about the ability to distinguish between medical school applicants in the near future unless A-levels are not reformed to enable the better performing students to be distinguished.

The paper looks at the A-level qualifications of all UCAS applicants, medical school applicants and medical school entrants. The authors find that some 15% of all applicants, 45% of medical school applicants and 62% of medical school entrants have the maximum 3As in their best 3 A-levels (excluding General Studies). The authors argue that, with nearly half of all medical school applicants getting the maximum score, selection becomes difficult. Fitting a censored normal distribution to the data, the authors estimate that only around 11% of medical school applicants would have scored at the level of 3 A*, with A* being the new A-level grade being introduced for students starting their A-levels in September 2008. However, the authors go onto to make the point that if grade
inflation continues at roughly the same rate as at present (over the last 10 years),
then by 2020 around one-third of all applicants would have 3 A*. The authors
therefore argue for the introduction for a A** classification at A-level to distinguish
truely outstanding performance at the projections suggest.

Response: This is a very fair summary of our paper.

-------------------------------------------------------------------------------

Major Compulsory Revisions (that the author must respond to before a decision
on publication can be reached)

The authors use a censored normal distribution to fit the distribution of A-level
performance of students. The fit of the censored normal to the data is shown in
Figures 1a, 1b, and 1c by comparing the black and yellow bars, which the
authors claim is "good", which is true except for the systematic underprediction
of 28 points and the overprediction of 30 points. Perhaps the authors ought to
consider some statistic on the appropriateness of the fit, e.g. a Pearson
goodness of fit measure.

Response: We have calculated a Pearson goodness of fit chi-square for
applicants and entrants scoring 12 points or more at A-level, which gives a value
of 1540.7 with 20 df. Inevitably, given the high sample size, this is statistically
very significant. This problem is of course well known in structural equation
modelling, where chi-square minimisation is a reasonable method of parameter
estimation, it is a poor method of goodness of fit testing since it is invariably
significant with large sample sizes. Various methods have been adopted to get
around this, one of which is Bollen's Incremental Fit Index (IFI) (see
www.ioa.pdx.edu/newsom/semclass/ho_fit.doc). Using any of these relative fit
methods requires some estimate of a null fit, and faute de mieux, we have used
equal frequencies at all A-level grades. That gives a chi-square statistic of
52194.4, so that IFI has a value of 0.971, which meets most standard criteria of
being greater than 0.95. We accept though that this is utterly arbitrary.

The over-estimation of the people on 30 points, will then lead to an
over-estimation of people projected to attain 36 points (3 A*) and 42 points (3
A**).

Response: This is a fair point, and we accept that our predictions are precisely
that B predictions B and that there may be better distributions for fitting the data
which avoid this problem. However, that does not undermine the main point of
the paper, which is that in the absence of any better assessment it is likely that
A* is not going to go on working for very long.

The approximation of the distribution of all applicants (when points range from 6
to 30) by a censored normal seems reasonable, but when the range of points is much narrower essentially 20-30 (so 6 distinct values for medical entrants) perhaps the approximation is less good.

Response: We agree, but to a large extent suspect that this is because with high A-level grades one is trying to estimate the bulk of a distribution using the tail alone, which will always have problems.

Perhaps the authors could show that their estimates of the proportions getting 3 A* or 3A** grades are reasonably robust to alternative approximations of the distributions, if it is the case that the goodness of fit tests reject the null.

Response: The problem here is that it is not at all clear what the alternative distributions are. As the reviewer comments, the set of data for all applicants is pretty well fitted by a normal distribution (with the minor exception of some over and under estimation of the AAA-AAB groups, for reasons that are not clear). It is therefore a reasonable assumption that any selected sub-group of candidates will also have a normal distribution, and it is that therefore that we used. There is an infinite number of non-normal distributions that could be used instead, but in the absence of any clear theoretical or empirical basis for distinguishing between them, we feel that for our immediate purpose that the normal distribution is adequate.

------------------------------------------------------------------------------------------------------------------------------------

Minor Essential Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct)

------------------------------------------------------------------------------------------------------------------------------------

Discretionary Revisions (which the author can choose to ignore)

What next?: Unable to decide on acceptance or rejection until the authors have responded to the major compulsory revisions

Level of interest: An article of importance in its field

Quality of written English: Acceptable

Statistical review: Yes, and I have assessed the statistics in my report.

Declaration of competing interests:
I declare that I have no competing interests

Responses to reviews. We are grateful to the reviewers for their thoughtful comments, and we hope we have responded adequately to them. In the following responses, for clarity, the original comments are in Roman script, and our responses are in italics.
Reviewer's report
Title: Even one star at Alevel could be "too little too late" for medical student selection.
Version: 1 Date: 16 September 2007
Reviewer: Sandra Nicholson
Reviewer's report:

General

I enjoyed reading the authors' analysis of the ability of current A Level grades to aid medical student selection with their cautionary prediction of how introducing A* grades will only facilitate discrimination by academic merit for approximately the next decade. Whilst previous work of the first author has demonstrated the predictive validity of success at A Level with medical school and postgraduate examinations I believe the authors were right to attempt to demonstrate again this assertion in the light of recent A Level inflation. The data concerning year 3 assessments (MCQ and OSCE) appear to corroborate with these previous findings. However we do not know, and the authors are assuming, that students with A*/A** will perform better in medical school examinations.

Response: This is a fair point, and reminds all educationalists that validity is not something that can be assumed for all tests for all times. Tests and times change, and validity should be routinely monitored for all selection methods. Having said that, until recently there was a surprising dearth of evidence that A levels were actually valid predictors of medical school outcome. We have added a brief comment on this in the discussion.

The authors are right to call for further research into the relationship between academic success as routinely measured by A Level grades and performance at medical school, but I would also add more importantly as practising doctors.

Response: We agree entirely with this and have added a line to the discussion saying, "Predictive validity should also be examined broadly, not merely for examination success, but for other aspects of real world clinical practice in working doctors."

This introduces my main philosophical point that whilst introducing A* and perhaps A** grades facilitates discrimination between applicants for medical courses, well demonstrated by this paper, in terms of who to reject it does not necessarily help select candidates on the basis of who will go on to be the best and happiest doctors. Surely there must be a ceiling on the relationship between academic success and performance at medical school and certainly with medical practice.
Response: Whether or not there is a ceiling to the predictive ability of A-levels or other educational measures is an empirical question and an important one. In unpublished data we have never found any evidence of non-linearity in the predictive validity of A-levels. Perhaps with the very large databases being collected by consortia such as UKCAT then the question will be properly unanswerable.

The paper does not acknowledge that most medical schools take into consideration, some highly, other measures of suitability for medical training. To be fair A Levels remain the most heavily weighted selection tool currently but with their discussed limitations it is time to seek other reliable and valid methods of selecting tomorrow's doctors. Therefore whilst I agree with the authors A* may be "too little too late" I would also add "too narrow" and urge the medical selection community to consider alternative selection tools alongside the traditional measurements of academic success.

Response: We accept this argument, and acknowledge fully that some medical schools do take a range of other measures into account. Having said that, good evidence of their validity is rare to the point of being almost entirely lacking. At present we think it is reasonable to assert that A-levels are the only selection measure in use in the UK that has proper evidence of long-term predictive validity. We will look forward to similar evidence being provided for other measures.

Major Compulsory Revisions (that the author must respond to before a decision on publication can be reached)
None

Minor Essential Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct)
On page 2
Results
The phrase "A* grades had been already introduced..." makes the meaning of the sentence unclear do they mean "if or had A* grades had been already introduced" ??

Response: We agree this is ambiguous and have changed the wording to, "If it were the case that A* grades had already been introduced, then...".

On page 3
Background: Bearing in mind my general comments above I would ask the authors to consider their use of "unvalidated tests" do they mean tests with no long term predictive validity which is fair comment if contextualised. It is not appropriate to simply say the newly introduced tests for medical student selection are unvalidated when there is evidence of construct validity and reliability amongst internal items and with similar tests in other fields.

Response: We accept that there is evidence of reliability and construct validity for the tests, and since submission of our original manuscript, BMAT has published (non-peer reviewed) data on its website suggesting some predictive validity for its test. We have therefore altered the paragraph to read:

"The result is that many medical schools are now using tests of general intellectual ability for selection for which until recently there was little evidence of predictive validity (although the assessments are reliable and have some construct validity) [1]. Since submission of the first draft of this paper, BMAT has provided some evidence for predictive validity for BMAT’s Section 2 (Scientific Knowledge and Applications) on academic examinations (see www.bmat.org.uk/background.html)."

Discretionary Revisions (which the author can choose to ignore)
On page 5
I would question whether the introduction of A*s was improving the process of selection. Making a process easier does not necessarily make it better and as indicated earlier this paper can, and perhaps should, indicate some of the competing complexities in student selection.

Response: Easier and better are not indeed equivalent, and we hope we have not implied that they are. However, we do believe that extending the range of an already valid measuring instrument to cover a wider range of candidates can only be an improvement.

Anecdotally some lesser able candidates sitting 4 A Levels may bring their average down. I wonder if the authors would like to comment on the reliability of their data for ALEVUCAS bearing this in mind?

Response: This is a fair point, which is why we only used the three best grades obtained by candidates sitting four A-level grades. There is a risk that candidates have diluted their effort by sitting four A levels when they would have done better by concentrating on three, but in practice there seems to be little evidence of this, presumably because teachers advise such students appropriately. We have never been happy with UCAS’ methods of analysing A-levels, but this is probably not an appropriate place to go further in that discussion.
What next?: Accept after minor essential revisions
Level of interest: An article of importance in its field
Quality of written English: Acceptable
Statistical review: Yes, but I do not feel adequately qualified to assess the statistics.
Declaration of competing interests:
I am a director of UKCAT and chair its test development working group. The UKCAT Consortium aim to develop a valid and reliable selection test for medical and dental degree programmes in the UK. These responsibilities do not involve any personal financial gain.

Reviewer's report
Title: Even one star at A level could be "too little too late" for medical student selection.
Version: 1 Date: 21 September 2007
Reviewer: Trudie T Roberts
Reviewer's report:
This paper is interesting and well written as one would expect of these authors. They make the assumption of constant levels of grade inflation. The work by Robert Coe et al suggests that this might be an over estimate. Never the less this paper is well argued in support of greater grade discrimination at the top end of the distribution and as far as I can tell the models used are appropriate. Although, it should be remembered that I am not a statistician.

Response: In August 2007 the Times reported that Robert Coe had suggested that grade inflation might be different in different subjects, and might be particularly high in maths (a subjects taken by many medical school applicants). As yet we have been unable to find a copy of this report, but are trying.

What next?: Accept without revision
Level of interest: An article of importance in its field
Quality of written English: Acceptable
Statistical review: Yes, but I do not feel adequately qualified to assess the statistics.
Declaration of competing interests:
I declare that I have no competing interests
Reviewer's report
Title: Even one star at Alevel could be "too little too late" for medical student selection.
Version: 1 Date: 16 November 2007
Reviewer: Jeremy Smith
Reviewer's report:
General
This was a well written and informative article, which raises questions about the ability to distinguish between medical school applicants in the near future unless A-levels are not reformed to enable the better performing students to be distinguished.

The paper looks at the A-level qualifications of all UCAS applicants, medical school applicants and medical school entrants. The authors find that some 15% of all applicants, 45% of medical school applicants and 62% of medical school entrants have the maximum 3As in their best 3 A-levels (excluding General Studies). The authors argue that, with nearly half of all medical school applicants getting the maximum score, selection becomes difficult. Fitting a censored normal distribution to the data, the authors estimate that only around 11% of medical school applicants would have scored at the level of 3 A*, with A* being the new A-level grade being introduced for students starting their A-levels in September 2008. However, the authors go on to make the point that if grade inflation continues at roughly the same rate as at present (over the last 10 years), then by 2020 around one-third of all applicants would have 3 A*. The authors therefore argue for the introduction for a A** classification at A-level to distinguish truly outstanding performance at the projections suggest.

Response: This is a very fair summary of our paper.

-------------------------------------------------------------------------------------------------

Major Compulsory Revisions (that the author must respond to before a decision on publication can be reached)

The authors use a censored normal distribution to fit the distribution of A-level performance of students. The fit of the censored normal to the data is shown in Figures 1a, 1b, and 1c by comparing the black and yellow bars, which the authors claim is "good", which is true except for the systematic underprediction
of 28 points and the overprediction of 30 points. Perhaps the authors ought to consider some statistic on the appropriateness of the fit, e.g. a Pearson goodness of fit measure.

Response: We have calculated a Pearson goodness of fit chi-square for applicants and entrants scoring 12 points or more at A-level, which gives a value of 1540.7 with 20 df. Inevitably, given the high sample size, this is statistically very significant. This problem is of course well known in structural equation modelling, where chi-square minimisation is a reasonable method of parameter estimation, it is a poor method of goodness of fit testing since it is invariably significant with large sample sizes. Various methods have been adopted to get around this, one of which is Bollen's Incremental Fit Index (IFI) (see www.ioa.pdx.edu/newsom/semclass/ho_fit.doc). Using any of these relative fit methods requires some estimate of a null fit, and faute de mieux, we have used equal frequencies at all A-level grades. That gives a chi-square statistic of 52194.4, so that IFI has a value of 0.971, which meets most standard criteria of being greater than 0.95. We accept though that this is utterly arbitrary.

The over-estimation of the people on 30 points, will then lead to an over-estimation of people projected to attain 36 points (3 A*) and 42 points (3 A**).

Response: This is a fair point, and we accept that our predictions are precisely that B predictions B and that there may be better distributions for fitting the data which avoid this problem. However, that does not undermine the main point of the paper, which is that in the absence of any better assessment it is likely that A* is not going to go on working for very long.

The approximation of the distribution of all applicants (when points range from 6 to 30) by a censored normal seems reasonable, but when the range of points is much narrower essentially 20-30 (so 6 distinct values for medical entrants) perhaps the approximation is less good.

Response: We agree, but to a large extent suspect that this is because with high A-level grades one is trying to estimate the bulk of a distribution using the tail alone, which will always have problems.

Perhaps the authors could show that their estimates of the proportions getting 3 A* or 3A** grades are reasonably robust to alternative approximations of the distributions, if it is the case that the goodness of fit tests reject the null.

Response: The problem here is that it is not at all clear what the alternative distributions are. As the reviewer comments, the set of data for all applicants is pretty well fitted by a normal distribution (with the minor exception of some over and under estimation of the AAA-AAB groups, for reasons that are not clear). It is
therefore a reasonable assumption that any selected sub-group of candidates will also have a normal distribution, and it is that therefore that we used. There is an infinite number of non-normal distributions that could be used instead, but in the absence of any clear theoretical or empirical basis for distinguishing between them, we feel that for our immediate purpose that the normal distribution is adequate.

-----------------------------------------------------------------------------------

Minor Essential Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct)

-----------------------------------------------------------------------------------

Discretionary Revisions (which the author can choose to ignore)
What next?: Unable to decide on acceptance or rejection until the authors have responded to the major compulsory revisions
Level of interest: An article of importance in its field
Quality of written English: Acceptable
Statistical review: Yes, and I have assessed the statistics in my report.
Declaration of competing interests:
I declare that I have no competing interests