PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (Error! Hyperlink reference not valid.) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Does performance at medical school predict success at the Intercollegiate Membership of the Royal College of Surgeons (MRCS) examination? A retrospective cohort study
AUTHORS	Ellis, Ricky; Scrimgeour, Duncan; Brennan, Peter; Lee, Amanda; Cleland, Jennifer

VERSION 1 – REVIEW

REVIEWER	Tiffin, Paul University of York, Health Sciences As a member of the UK medical education database (UKMED) research subgroup I reviewed the original application to UIKMED for data access for this project, and also the subsequent report submitted to the UKMED research group.
REVIEW RETURNED	15-Dec-2020

GENERAL COMMENTS	This is an interesting and relatively concise paper that adds usefully to the literature. I hope that it will be published, with appropriate revisions, fairly promptly, as it could contribute informatively to the current discussions regarding the foundation programme allocation in the UK. Abstract and Introduction I only have a few minor comments about this. It is probably slightly more precise to talk about an increase of 52% in the 'odds', rather than 'chances' of passing the MRCS at first attempt. Also, although I make further comments in the methods section, independent prediction here, should probably mean one of two things, to place the findings properly in the policy context. Firstly, it should either mean 'independent of other educational or performance metrics used in the foundation program selection', if that is to be the focus? Secondly, it might mean 'independence of other potential confounding variables, such as demographic factors'. Although generally we do not select these latter variables there may be implications for understanding how these measures work, but also the influence on the demographics and characteristics of the final population selected (or 'allocated' in the case of the FP). In contrast, the former issue indicates the incremental validity of a selection measure. Unless these two issues are separated clearly the policy implications for the multivariable regression results are not clear. Overall, the introduction was well written and concise. Although the concept of the 'academic backbone' is mentioned in the

	discussion, it would probably be useful to mention it in the introduction also, as it sets the context. Also, it would be good to be clearer about what the SJT for FP allocation is likely to measure. Such SJTs don't actually directly "test… the behaviours and attitudes expected of doctors as described in the General Medical Council's" as stated in the introduction but might be better considered as a special case of a knowledge test; specifically "…such knowledge would include procedural knowledge about what to do in certain situations and how to do it. (Tiffin, Paton, O'Mara et al. 2020 Med Ed). Methods Methods seem largely appropriate. However, in my view the stepwise model building was not the optimum way of conducting the multivariable analyses (overlooking more general criticisms of stepwise approaches). This is because as, as highlighted earlier, the independent effects need to be applicable to the policy questions in hand. That is, to what degree, if any, do the different elements of FP allocation add incremental (independent) contributions? Thus it is important to separate off these two analyses so that the policy implications can be informatively discussed. Thus, findings from at least two set of multivariable analyses should reported: those where all the FP allocation assessment components are entered into the model, and; one where any relevant demographic or educational factors are entered (e.g. graduate status, gender etc). Moreover, there should probably be a third set of multivariable analyses that look at whether any association between MRCS performance and the FP selection measure is explained by the demographic factors. Being familiar with the UKMED data I don't think there would be extensive missing data would be present, but this should be explicit, and how any missing values were handled. Also, in line with STROBE guidelines there should be a figure showing the flow of data through the study, and what the level of completeness was for the multivariable analyses, especially if listwise deletion was used in terms of missingness handling. I was pleased interactions were evaluated though there are many schools of thought about to what degree all permutations of interaction terms should be explored, especially as some resulting coefficients may be difficult to interpret. Also- I could not see if the raw MRCS scores were used or scores 'relative to pass'? We always use the latter in our research into postgraduate performance as raw scores are more likely to vary in terms of meaning across diets. Also- UKMED research is covered by a blanket ethical approval and this should be stated in the manuscript. Results The usual table summarising the demographics of the analytic sample is missing though probably most of the information can still be gleaned from Table 2, which appears to be a reasonable substitute. I am not sure about using Pearson's moment correlation for relations with EPM, which should have, I think, a rectangular distribution (being based on ranking), though I suspect the findings would be similar if Spearman's rho, as a non-

	parametric alternative is employed. Otherwise, these results are straightforward and presented relatively clearly, though as stated earlier the multivariable models for selection measure vs demographics etc should be analyses and reported separately. Also, Table legends should be clearer about which are the univariable and multivariable results and preferably have an associated p value in a separate column as well as the CIs (I know the p value is in disrepute but I still think it flags up statistically significant findings helpfully). Discussion The discussion could be developed quite a bit further, with perhaps a clearer message for policy. Especially if the multivariable analyses are re-run as suggested. Also- our recently published systematic review and meta-analysis is probably the best reference to cite for the validity of SJTs in medical selection (Webster et al. 2020 Med Ed). We did actually include information on the FP SJT in this recent evidence synthesis. As mentioned earlier, a slightly different approach to the multivariable analyses would have fed and enriched the discussion (ie. being clearer about how the predictive validity of each FP allocation element independent of the other components. A key limitation of studies using the EPM is that the EPM is a local not a national measure. At present most Medical Schools in the UK are fairly similar to each other though I am unsure whether this trend will continue. I developed a method to 'nationalise' the EPM (using something called 'peer competition rescaling'. However, it didn't seem to change the correlation between EPM and national measures (e.g. MRCP scores) substantially, which is reassuring. Nevertheless the local, ranking, nature of EPM should be emphasised as a potential limitation to this study. The authors are correct to raise the issue of the MLA, and whether it is likely to add value to medical selection and regulation. The issue of (differential) attainment is also mentioned in the discussion. I am not sure whether this should be the subject of a separate paper? We recently published a study, purely focussed on this issue in relation to MRCPsych (Tiffin & Paton, 2020, PMJ) and it included quite a number of analyses to try and better understand the issues in this postgraduate clinical education context. I am tempted to recommend that this paper purely focuses on the FP allocation measures, and a separate report could explore the important, and sensitive issue of differential attainment, so as to more thoroughly do it justice? My view is that the demographics should be included in the analyses but only in relation to whether any association between any association between MRCS performance and the FP selection measure is explained by the demographic factors. This could shed light on how any associations with the selectin (allocation) measures are mediated, but would keep the study focus on the FP allocation process, rather than straying into the differential attainment area. Otherwise, I think this paper is a useful contribution to the literature I would be delighted to review a revised version, if appropriate.

REVIEWER	Mathews, Maria Memorial University of Newfoundland, Community Health and Humanities
REVIEW RETURNED	11-Jan-2021

GENERAL COMMENTS	The paper provides a good background to Education performance Measures (EPM) and the membership of the Royal College of Surgeons (MRCS) examination (Part A and Part B). Variables in the study should be consistently defined and treated in the analysis as either as categorical or continuous variables. For example, graduate status and publications are dichotomized into yes/no variables; but Table 3 presents means and standard deviations for them. Similarly, exam outcomes are dichotomized into passed (yes/no) but Pearson correlations were done with selection scores and MRCS Part A in Table 4. In table 5, graduate status appears twice, coded as both a continuous variable and categorical variable Table 2 suggests there are two samples: MRCS part A (n= 1975) and MRCS Part B (n=630). A clearer description of how each sample was derived is needed. Table 2 is confusing. I suggest that the authors reorganize the table to summarize the characteristics of the two samples (i.e. present frequencies) and create additional tables to compare the characteristics of students who passed and did not pass Part A and Part B of the exam. I suggest they present column per cents. Gender and ethnicity variables are entered as an interaction term. It would be easier to understand if they were recoded into 4 groups (e.g. white male, non-white male, white female and non-white female) throughout the analysis. Were additional variables such as year or medical school/region available? Discussion covers relevant findings and highlights study strengths and limitations.

REVIEWER	Sen Gupta, Tarun James Cook Univ
REVIEW RETURNED	19-Jan-2021

GENERAL COMMENTS	This is a well-written paper that adds further evidence to the accepted wisdom that 'past results predict future results.' Given the large dataset and link between undergraduate and postgraduate education it is likely to be of broad interest. My main feedback relates to the significance of assessing success by examination results. Many authors have argued we should 'count what counts' (see for example DOI: 10.31128/AJGP-02-18-4488). In other words, a threshold ability is important(ie competence, or passing exams) but beyond that the social accountability literature argues that it is what graduates do with their degree, and the communities they serve that is important.

	Being able to use prior results to predict, to some degree, postgraduate results can be used to set a threshold eg those likely to succeed. This does not necessarily mean selecting the top students, but perhaps a broader pool, also selected for other attributes like gender, ethnicity, rurality and other underserved minorities. Some of these data were available to the investigators, so I would suggest further considering this point, or making a statement about why it was not considered). Some minor points for consideration: - the term 'EPM decile point' (eg in Abstract / Results) took a little time to digest; I suggest this is explained in a little more detail, and the reasons for choice of this methodology justified - I could not see the source data to explain the increased pass rate relating to the deciles. Including this would be helpful; also suggest making clear this is a relative increase, not absolute. - I am not sure that the EPM actually is an example of programmatic assessment (p7/33) - suggest this be clarified or a reference provided - Table 2 seems to have an error in the first row of data: the numbers in part A (1490 and 480) add to 1970, not 1975 - The line starting 'Whilst' (p15/33) needs a verb Thank you for the opportunity to review this substantial body of work.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1: This is an interesting and relatively concise paper that adds usefully to the literature. I hope that it will be published, with appropriate revisions, fairly promptly, as it could contribute informatively to the current discussions regarding the foundation programme allocation in the UK. Abstract and Introduction I only have a few minor comments about this. It is probably slightly more precise to talk about an increase of 52% in the 'odds', rather than 'chances' of passing the MRCS at first attempt.	This has now been clarified.
Also, although I make further comments in the methods section, independent prediction here, should probably mean one of two things, to place the findings properly in the policy context. Firstly, it should either mean 'independent of other educational or performance metrics used in the foundation program selection', if that is to be the focus? Secondly, it might mean 'independence of other potential confounding variables, such as demographic factors'. Although generally we do not select these latter variables there may be implications for understanding how these measures work, but also the influence on the	Now that the multivariate analysis has been adjusted as per the reviewer's suggestions, the first column in Table 5 shows predictors of success independent of other performance metrics and the second column show the predictors of both other performance metrics and demographic factors.

demographics and characteristics of the final population selected (or 'allocated' in the case of the FP). In contrast, the former issue indicates the incremental validity of a selection measure. Unless these two issues are separated clearly the policy implications for the multivariable regression results are not clear.	
Overall, the introduction was well written and concise. Although the concept of the 'academic backbone' is mentioned in the discussion, it would probably be useful to mention it in the introduction also, as it sets the context.	This have now been mentioned.
Also, it would be good to be clearer about what the SJT for FP allocation is likely to measure. Such SJTs don't actually directly "test… the behaviours and attitudes expected of doctors as described in the General Medical Council's" as stated in the introduction but might be better considered as a special case of a knowledge test; specifically "…such knowledge would include procedural knowledge about what to do in certain situations and how to do it. (Tiffin, Paton, O'Mara et al. 2020 Med Ed).	This have now been changed.
Methods Methods seem largely appropriate. However, in my view the stepwise model building was not the optimum way of conducting the multivariable analyses (overlooking more general criticisms of stepwise approaches). This is because as, as highlighted earlier, the independent effects need to be applicable to the policy questions in hand. That is, to what degree, if any, do the different elements of FP allocation add incremental (independent) contributions? Thus it is important to separate off these two analyses so that the policy implications can be informatively discussed. Thus, findings from at least two set of multivariable analyses should reported: those where all the FP allocation assessment components are entered into the model, and; one where any relevant demographic or educational factors are entered (e.g. graduate status, gender etc).	Multivariate analyses in Table 5 have been adjusted to meet the reviewers' recommendations. Two sets of analyses are now presented as suggested.
Moreover, there should probably be a third set of multivariable analyses that look at whether any association between MRCS performance and the FP selection measure is explained by the demographic factors.	As suggested by the reviewer, we have now included performance metrics alone before adjusting for sociodemographic factors. The focus of the paper is to establish whether performance at medical school predicts MRCS performance not whether sociodemographic differences predict MRCS performance – this is already known (Scrimgeour et al, 2018

	https://doi.org/10.1016/j.surge.2017.10.001)). Therefore, having discussed this with all the authors, we do not think that a further analysis of demographic factors alone is relevant to our specific research questions here. In addition, we have also reduced the discussion around differential attainment as per the reviewers' later suggestion.
Being familiar with the UKMED data I don't think there would be extensive missing data would be present, but this should be explicit, and how any missing values were handled. Also, in line with STROBE guidelines there should be a figure showing the flow of data through the study, and what the level of completeness was for the multivariable analyses, especially if listwise deletion was used in terms of missingness handling. I was pleased interactions were evaluated though there are many schools of thought about to what degree all permutations of interaction terms should be explored, especially as some resulting coefficients may be difficult to interpret.	We have added a data flow diagram as suggested.
Also- I could not see if the raw MRCS scores were used or scores 'relative to pass'? We always use the latter in our research into postgraduate performance as raw scores are more likely to vary in terms of meaning across diets.	Scores relative to pass were used in this study. This has now been clarified in the methods section.
Also- UKMED research is covered by a blanket ethical approval and this should be stated in the manuscript.	This has now been added.
Results The usual table summarising the demographics of the analytic sample is missing though probably most of the information can still be gleaned from Table 2, which appears to be a reasonable substitute.	We have now expanded table 2 to include all demographics and frequencies of the study population.
I am not sure about using Pearson's moment correlation for relations with EPM, which should have, I think, a rectangular distribution (being based on ranking), though I suspect the findings would be similar if Spearman's rho, as a non-parametric alternative is employed.	We now use Spearman's Rho for EPM Decile.
Otherwise, these results are straightforward and presented relatively clearly, though as stated earlier the multivariable models for selection measure vs demographics etc should be analyses and reported separately.	Data analyses have now been extended and adapted to meet the reviewer's suggestions.
Also, Table legends should be clearer about which	This has now been clarified in the methods

are the univariable and multivariable results and preferably have an associated p value in a separate column as well as the Cis (I know the p value is in disrepute but I still think it flags up statistically significant findings helpfully).	section.
Discussion The discussion could be developed quite a bit further, with perhaps a clearer message for policy. Especially if the multivariable analyses are re-run as suggested.	
Also- our recently published systematic review and meta-analysis is probably the best reference to cite for the validity of SJTs in medical selection (Webster et al. 2020 Med Ed). We did actually include information on the FP SJT in this recent evidence synthesis.	This paper has now been referenced.
As mentioned earlier, a slightly different approach to the multivariable analyses would have fed and enriched the discussion (ie. being clearer about how the predictive validity of each FP allocation element independent of the other components.	The discussion has been revised and now focusses on the incremental value of FP selection tools and implications for policy.
A key limitation of studies using the EPM is that the EPM is a local not a national measure. At present most Medical Schools in the UK are fairly similar to each other though I am unsure whether this trend will continue. I developed a method to 'nationalise' the EPM (using something called 'peer competition rescaling'. However, it didn't seem to change the correlation between EPM and national measures (e.g. MRCP scores) substantially, which is reassuring. Nevertheless the local, ranking, nature of EPM should be emphasised as a potential limitation to this study. The authors are correct to raise the issue of the MLA, and whether it is likely to add value to medical selection and regulation.	This has now been addressed in the discussion.
The issue of (differential) attainment is also mentioned in the discussion. I am not sure whether this should be the subject of a separate paper? We recently published a study, purely focussed on this issue in relation to MRCPsych (Tiffin & Paton, 2020, PMJ) and it included quite a number of analyses to try and better understand the issues in this postgraduate clinical education context. I am tempted to recommend that this paper purely focuses on the FP allocation measures, and a separate report could explore the important, and sensitive issue of differential attainment, so as to more thoroughly do it justice? My view is that the demographics should be	Thank you for your recommendations. We agree and have removed the discussion regarding differential attainment. This will now be explored in a separate paper. The discussion focusses on the implications of these results on policy decisions has been extended.

included in the analyses but only in relation to whether any association between any association between MRCS performance and the FP selection measure is explained by the demographic factors. This could shed light on how any associations with the selectin (allocation) measures are mediated, but would keep the study focus on the FP allocation process, rather than straying into the differential attainment area. Otherwise, I think this paper is a useful contribution to the literature I would be delighted to review a revised version, if appropriate.	
Reviewer: 2 The paper provides a good background to Education performance Measures (EPM) and the membership of the Royal College of Surgeons (MRCS) examination (Part A and Part B). Variables in the study should be consistently defined and treated in the analysis as either as categorical or continuous variables. For example, graduate status and publications are dichotomized into yes/no variables; but Table 3 presents means and standard deviations for them. Similarly, exam outcomes are dichotomized into passed (yes/no) but Pearson correlations were done with selection scores and MRCS Part A in Table 4. In table 5, graduate status appears twice, coded as both a continuous variable and categorical variable	It appears that different variables may have been misinterpreted by the reviewer. It states in the Methods section that “Except for SJT and EPM scores, all variables were subsequently dichotomized”. EPM scores including degree and publication scores remained continuous variables throughout the analyses. Being a graduate (which was dichotomised as yes/no) is not the same as achieving EPM degree points. For example, intercalating whilst studying medicine as an undergraduate would earn EPM degree points, but the student still entered medicine without a prior degree, classifying them as an undergraduate. Similarly, more EPM degree points are awarded for the grade achieved on a degree, not for the number or undergraduate degrees completed.
Table 2 suggests there are two samples: MRCS part A (n= 1975) and MRCS Part B (n=630). A clearer description of how each sample was derived is needed. Table 2 is confusing. I suggest that the authors reorganize the table to summarize the characteristics of the two samples (i.e. present frequencies) and create additional tables to compare the characteristics of students who passed and did not pass Part A and Part B of the exam. I suggest they present column per cents.	To clarify this, we have added a data flow chart (Figure 1) and have added new frequencies to the revised Table 2.
Gender and ethnicity variables are entered as an interaction term. It would be easier to understand if they were recoded into 4 groups (e.g. white male, non-white male, white female and non-white female) throughout the analysis.	We thank the reviewers for their valuable comments and suggestions, but in order for this paper to be applicable for policy decisions we feel that it is important to bring it in line with other large UKMED studies that have analysed results in this way. Additionally, subcategorising these variables into four groups would not clarify the interaction term, but would instead add further complexity since one group would need to be defined as a reference category. Lastly, as per the suggestions of reviewer 1, we have removed discussion of

	differential attainment which will need further careful exploration and discussion in a separate paper.
Were additional variables such as year or medical school/region available?	These will be addressed in a separate paper.
Reviewer: 3 This is a well-written paper that adds further evidence to the accepted wisdom that 'past results predict future results.' Given the large dataset and link between undergraduate and postgraduate education it is likely to be of broad interest. My main feedback relates to the significance of assessing success by examination results. Many authors have argued we should 'count what counts' (see for example DOI: 10.31128/AJGP-02-18-4488). In other words, a threshold ability is important (ie competence, or passing exams) but beyond that the social accountability literature argues that it is what graduates do with their degree, and the communities they serve that is important. Being able to use prior results to predict, to some degree, postgraduate results can be used to set a threshold eg those likely to succeed. This does not necessarily mean selecting the top students, but perhps a broader pool, also selected for other attributes like gender, ethnicity, rurality and other underserved minorities. Some of these data were available to the investigators, so I would suggest further considering this point, or making a statement about why it was not considered).	We completely agree with this statement, and indeed this point has been made by co-author Professor Jen Cleland in many papers. However, the focus of this paper is predictive validity, rather than social accountability. We also think this is a significant point of discussion, one which merits more than a comment and so is best left for another paper. In saying this, we believe our extended discussion around the value of the EA makes a pertinent point re what is important to count.
Some minor points for consideration: the term 'EPM decile point' (eg in Abstract / Results) took a little time to digest; I suggest this is explained in a little more detail, and the reasons for choice of this methodology justified	We acknowledge that the term ‘EPM decile point’ is a little confusing to those not familiar with the EPM or foundation selection methods, but it is the most accurate description of the performance metric. We have therefore, explained the meaning of this term and how candidates are awarded points based on their deciles on page 5 of the manuscript.
I could not see the source data to explain the increased pass rate relating to the deciles. Including this would be helpful; also suggest making clear this is a relative increase, not absolute.	Thank you for your suggestion. Figure 2 has now been added to the manuscript to address this.
I am not sure that the EPM actually is an example of programmatic assessment (p7/33) - suggest this be clarified or a reference provided	This has now been clarified.

Table 2 seems to have an error in the first row of data: the numbers in part A (1490 and 480) add to 1970, not 1975	This typographical error has been corrected.
The line starting 'Whilst' (p15/33) needs a verb	This has now been changed.

VERSION 2 – REVIEW

REVIEWER	Sen Gupta, Tarun James Cook Univ
REVIEW RETURNED	10-Jun-2021

GENERAL COMMENTS	Thank you for the opportunity to again review this paper and for addressing the comments from the first review round so comprehensively. I believe the paper is now much stronger, and have only some minor suggestions for the authors to consider: - I note the previous reviewer's comment about "an increase of 52% in the 'odds' rather than the 'chances'.." which has now been addressed. However, I still struggled a little with statements like "For every additional EPM decile point gained, the odds of passing MRCS increased by 55% for Part A" and suggest an additional line is inserted to explain this eg 'OR for Decile 1 is X, and for Decile 2 is Y...' - the sentence on p6 starting 'The EPM is calculated out of 50 points and comprises three parts (Table 1)' is long, and hard to read. I would suggest reformatting (eg using 3 dot points) or revising the punctuation. The mention of "medical school performance decile (points are awarded depending on a student's final EPM decile..." seems to suggest that one component of the EPM is the EPM, which is circular and may be an error. I note the response to my point about social accountability, and the mention of the need to widen access to medical education, thank you.

VERSION 2 – AUTHOR RESPONSE

Comments to the Author: Thank you for the opportunity to again review this paper and for addressing the comments from the first review round so comprehensively. I believe the paper is now much stronger, and have only some minor suggestions for the authors to consider: - I note the previous reviewer's comment about "an increase of 52% in the 'odds' rather than the	We appreciate that communicating the meaning of odds ratios in this type of study is quite difficult to do in a way that is accurate yet straightforward for readers to interpret. However, we have adjusting our wording a little to improve "readability".

'chances'.." which has now been addressed. However, I still struggled a little with statements like "For every additional EPM decile point gained, the odds of passing MRCS increased by 55% for Part A" and suggest an additional line is inserted to explain this eg 'OR for Decile 1 is X, and for Decile 2 is Y...'	
- the sentence on p6 starting 'The EPM is calculated out of 50 points and comprises three parts (Table 1)' is long, and hard to read. I would suggest reformatting (eg using 3 dot points) or revising the punctuation. The mention of "medical school performance decile (points are awarded depending on a student's final EPM decile..." seems to suggest that one component of the EPM is the EPM, which is circular and may be an error. I note the response to my point about social accountability, and the mention of the need to widen access to medical education, thank you.	This has now been clarified, thank you.