The Relative Weighting of the SJT, EPM and Additional Educational Points In Selection for Foundation

Background

There are currently two research papers relating to the predictive abilities of the SJT and the EPM, and it is important to consider both of them together[i],[ii].

Smith and Tiffin studied 2 cohorts of applicants to the Foundation programme, dating from 2013 and 2014, and numbering 14,131 in total. The key outcome measure was performance at ARCP, with data available through the UK Medical Education Database (UKMED)[iii].

Cousans *et al* also studied 2013 applicants. The top and bottom two deciles of SJT and EPM scores were considered. Outcome data included subsequent personal evaluations of named Foundation trainees in practice by their educational supervisors, and scoring of the likelihood of necessary remedial action.

Tiffin *et al* observed that both the EPM and the SJT decile scores predicted the likelihood of successfully completing the programme as measured by obtaining ARCP outcome 6. When scores were attempted to be converted to a common scale, there was a difference between the EPM z-score decile (OR 1.14, 95% CI 1.10 to 1.18, p<0.001) and SJT z-score decile (OR 1.05, 95% CI 1.01 to 1.09, p=0.02).

Cousans *et al* observed that both SJT and EPM SJT scores correlated with supervisor ratings (*r* = 0.31 and 0.28, respectively). The relationship was stronger between the SJT and in-role performance for the low scoring group (*r* = 0.33, high scoring group *r* = 0.11), and between academic performance and in-role performance for the high scoring group (*r* =0 .29, low scoring group *r* =0 .11). Trainees with low SJT scores were almost five times more likely to receive remedial action.

It can be seen that the outcome measures are very different. The Cousans *et al* measures are based on observation of candidates and hence are subjective in nature. Since they are based on direct observation, they are many fewer in number than in Smith and Tiffin. The Smith and Tiffin measures are also subjective, in that they are based on ARCP outcomes, whose validity is contested[iv], but on the other hand there are a great many of them.

On the basis of their respective findings, Cousans *et al* also conclude that both measures are useful, and that the SJT may be more sensitive at the lower end of the scale. They do not explicitly comment on relative weightings as far as selection for Foundation goes, but an implication can be drawn that they believe that these should be approximately equal (for instance, see comment that "It is notable that the effect size of the SJT and EPM's relationships with the outcome criteria is approximately equal in their appropriate range (*r* =0 .29 for the SJT, *r* =0 .33 for the EPM)".

Smith and Tiffin indicate that, viewed by decile, the EPM has a higher impact, and suggest that this be reflected in the relative weightings of the two in selection for Foundation, with the EPM being approximately twice as heavily weighted as the SJT. They also suggest that the additional points available for educational achievements are of no measurable value by their metrics.

Comments

What follows is my personal view.

The Cousans *et al* paper features fewer candidates, due to the personal observational nature of the outcome measures, but there is no evidence that the paper is underpowered. However, there may be an issue with selective responding on the part of educational supervisors.

The Smith and Tiffin paper data is left-censored, in that in 2013 and 2014, a number of very low scoring SJT candidates (those 4 SD below the mean) were removed from the allocation process (35 in 2013 and 33 in 2014). Only one of these was re-instated (in 2014). These were the bottom scoring candidates. It is not at all implausible that these candidates would have run into problems during Foundation: indeed, if they were not to do so, then the process of removing them from the allocation list should be questioned in its entirely. While the numbers of candidates removed in this way may be small, so is the number of candidates in the Smith and Tiffin paper who failed to progress as normal. I make this number 338 'non outcome 6' candidates over the two years under study. On this scale, 68 candidates may well make a significant difference.

There will also have been an unknown number of candidates who failed Finals, but scored above 4.0 SD below the mean. These would not have been present in the Smith and Tiffin analysis, but it is plausible to assume, on the evidence in both research papers, that they would have been relatively low scoring on the SJT, and of course, they would be below the lowest decile of the EPM.

The Smith and Tiffin paper originally found an approximately equal effect of SJTs and EPMs, and only on transformation of the data did a differential appear. The authors state: *"It is also worth commenting that, at first glance, both the EPM deciles and the SJT scores appear equally predictive of completion of the foundation programme. However, when we attempt to place both measures, although be [sic] crudely, on the same scale (ie, divided into deciles), it is clear that EPM deciles are more predictive of this outcome".*

I have a concern about this process (which is different to 'not liking the outcome'!). The EPM scores are already in deciles: this is a low information state. Converting to z-scores does not affect this: the data remains in deciles, with a different numeric value. However, the SJT data is normally distributed, and negatively skewed. If it is converted into deciles *and each data point is analysed by its decile rank rather than its original value*, then information will have been lost. It is not clear to me from the paper that this has not happened. Since the SJT is described as being most sensitive at the low end, this could be particularly significant.

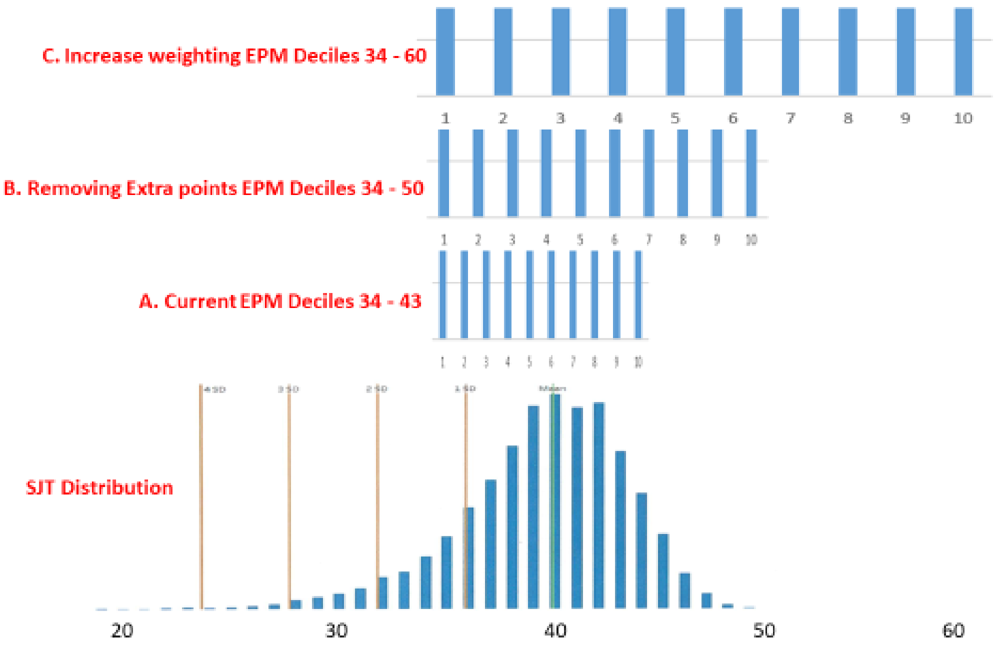
On the basis of the evidence currently available, I would class the relative predictive validity of the SJT and EPM as *contested*. Only further studies will enable a clear conclusion to be drawn.

Since administrative systems have inertia, and students will have employed long term strategies for optimising their outcomes, I would recommend that where data is validly debatable, the *status quo* should continue.

While the Cousans *et al* paper has little to say about the additional educational points, I would reiterate that students will have made long term commitments to gaining additional points, and might have a just grievance if the rules were changed without due notice. Since Intercalation frequently occurs between year 4 and 5 of a medical programme, two years' notice would be the minimum that should be given to dropping the additional points. But I understand that some medical schools permit Intercalation between years 2 and 3, and this should be explored in advance of any change.

There is a final important point which has not yet been considered in discussions of changing the relative weightings of the various components, and it is one that I think is very important.

It is that the range of 34-43 for the EPM deciles was chosen to correspond approximately to the mean of the SJT score plus or minus on standard deviation. If the range is changed there are major implications for the weighting which may not have been considered. In mathematical terms, the weighting is proportionate to the standard deviation[v],[vi]. Thinking it may be more intuitive to see this graphically, I've made the following image, which is not precise, but conveys the idea:



A shows approximately the current situation, where the decile range corresponds more-or-less to the most common scores on the SJT.

B shows the effect on the decile range of scoring the EPM from 34 to 50. The top two EPM deciles (20% of candidates) in particular represent a score very few SJT candidates can reach.

C shows the effect on the decile range of scoring the EPM from 34 to 60. Now the top 50% of EPM candidates massively outscore the great majority of SJT candidates.

The effect is exacerbated by the fact that the deciles have a flat distribution, rather than a normal distribution. And in this image, I've left the SJT at 50%, since I don't have the raw data to recalculate at 40 %. If the SJT were reduced to 40%, then the impact is even worse.

Any change to the relative weightings of the components must take this effect into account.

CoI

Note that I have potential conflicts of interest with regard to both papers. I am a former colleague of, and co-author with, Dr Tiffin, and serve as a member of his NIHR Fellowship Steering Group: and rely on the support of Daniel Smith for our current UKMED project. I was invited to review the Cousan's *et al* paper in draft, made some suggestions, and was subsequently invited to be included in the authorship. These CoI exist in the academic, not the financial, realm.

John C. McLachlan,

10th February, 2020

References

[i] Smith, D.T. and Tiffin, P.A., 2018. Evaluating the validity of the selection measures used for the UK's foundation medical training programme: a national cohort study. *BMJ open*, *8*(7), p.e021918.

[ii] Cousans, F., Patterson, F., Edwards, H., Walker, K., McLachlan, J.C. and Good, D., 2017. Evaluating the complementary roles of an SJT and academic assessment for entry into clinical practice. *Advances in Health Sciences Education*, *22*(2), pp.401-413.

[iii] https://www.ukmed.ac.uk/

[iv] Viney, R., Rich, A., Needleman, S., Griffin, A. and Woolf, K., 2017. The validity of the Annual Review of Competence Progression: a qualitative interview study of the perceptions of junior doctors and their trainers. *Journal of the Royal Society of Medicine*, *110*(3), pp.110-117.

[v] Eva, K.W. and Reiter, H.I., 2004. Where judgement fails: pitfalls in the selection process for medical personnel. *Advances in Health Sciences Education*, *9*(2), pp.161-174.

[vi] McLachlan, J.C. and Whiten, S.C., 2000. Marks, scores and grades: scaling and aggregating student assessment outcomes. *Medical education*, *34*(10), pp.788-797.