Analysis of Confidence Rating Pilot Data: Executive Summary for the UKCAT Board

Paul Tiffin & Lewis Paton

University of York

Background

Self-confidence may be the best 'non-cognitive' predictor of future academic performance (Stankov, Morony et al. 2014). Moreover, it is important that future doctors are neither over, nor under-confident in their own abilities if they are to practice medicine safely and effectively. Consequently, as outlined in the initial report by Pearson VUE (Pearson VUE 2017) from 2013 to 2016 a confidence rating was piloted within the Decision Analysis (DA) section of the cognitive section of the UKCAT. After each DA item, candidates were asked to use a scale from 1 to 5 to indicate how confident they were that the answer they gave was correct.

We analysed these data to explore whether 'online confidence', estimated by two different methods, was predictive of either the odds of success at application, or academic performance during the first two years of medical school. It should be noted that '*online confidence'* (confidence in your ability to answer online test items correctly) is conceptualised as somewhat distinct from self-reported confidence generally, which is better considered as a personality trait. In contrast 'online confidence' is probably best conceptualised as a 'meta-cognitive' attribute- i.e. one which requires a cognitive judgment about one's cognitive performance.

Methods

Confidence was estimated both using the conventional method ('*confidence bias*') used by Pearson VUE in the original pilot report, and by a novel method that sought to control for a number of potential sources of bias ('*confidence judgment*'). The *confidence bias* (CB) *score* was defined as the average difference between the confidence score and the accuracy score on the DA cognitive items. In order to calculate the confidence score for the UKCAT candidates, the confidence ratings ('1' to '5') were converted to 'confidence scores' (0, 0.25, 0.5, 0.75, or 1). The *confidence bias* was then calculated for each item by subtracting the candidate confidence score on the item from their actual score (0 or 1) achieved on the item. The confidence bias score for each candidate was therefore the mean confidence bias score for all the operational items that the candidate responded to.

In order to address potential sources of bias a novel approach to modelling online confidence was developed. In particular, as the self-ratings of confidence (one to five on a Likert scale) were not linked to any descriptive anchor points, individual candidates may have interpreted these points differently. Moreover, candidates may have had varying tendencies to use the extreme, or middle points when answering (i.e. 'extreme responders', versus 'the central tendency'). For this reason the effects of such 'lift and scatter' were controlled for by producing 'internal Z scores' for each item per candidate (McDaniel, Psotka et al. 2011). That is, the scores were standardised by subtracting the mean scores and dividing by the standard deviation, within each set of respondent's responses.

Secondly, in the conventional method the *confidence bias* score was generated by subtracting the 'normalised' (i.e. 0.00 to 1.00 for 1-5 ratings) perceived confidence from the

actual achieved score (1 or 0). However, using this approach, where candidates were generally highly confident (i.e. selecting mostly four or fives) but regularly obtaining incorrect answers there may have been a disproportionate effect, generating highly negative confidence scores for even a few incorrect responses. The impact of this can be softened by using item response theory to model the *expectation* (or in this case log odds) that a candidate, of a given ability, would have got that particular question right. In this instance a two parameter logistic (2-PL) regression IRT model was used to predict the probability (specifically the 'log odds') of a particular candidate getting a specific item correct. That is, the probability of a candidate of ability θ_j is observed obtaining a correct answer to a specific item (i.e. that the outcome, Y=1) is a function of the difference between the candidate's ability (θ_j) and the item difficulty (b_i), weighted by item discrimination (a_i- that is the ability of the item to differentiate candidates of different ability levels). Thus, the expectation that a respondent will answer a specific answer correctly is estimated by considering the difference in the ability of the individual and the difficulty of the question, whilst making allowances for the fact that some items will be better at discriminating between candidates than others.

Having obtained individual Z scores for each candidate's perception of their confidence in each specific item, as well as deriving an estimate of the probability that a particular candidate would be expected to get a particular item correct, given their overall ability level we were then able to examine the relationship between the two. If candidates are accurate at appraising their own abilities there should be a reasonable to strong correlation between their relative confidence for each item and the estimate that the candidate would be expected to get such items correct. Candidates who have limited ability to appraise their own abilities at such a cognitive test would be expected to show little or no correlation between these variables. Overconfident candidates may even show a negative correlation between these two variables. Thus, we were able to derive a correlation coefficient, representing the correlation between a candidate's relative confidence on an item and the estimate from the IRT model regarding the likelihood the candidate would have obtained a correct answer for that specific item. This required the generation of individual correlation coefficients for every candidate in the dataset, repeated for each form of the test (the IRT model varied according to the subset of items administered (per test form)).

Having derived estimates of both 'confidence bias' and 'confidence judgment' we were able to evaluate their association with both sociodemographic characteristics of candidates as well as performance at application and following entry to medical school (where successful).

Results

In terms of 'confidence bias', a tendency to *under-confidence* positively correlated with cognitive ability. This effect is probably partly mediated by the fact that those who tend to score 1s (i.e. obtain correct answers) only had to rate their confidence is less than 5 on a few items in order to get a positive *confidence bias* score.

Relatively lower *confidence bias* scores (overconfidence) were significantly ($p<0.0001$) associated with male sex, English as a Second language (EASL), declared non-White ethnicity and attending a non-selective secondary school. For these background variables we also controlled for the ability at *Decision Analysis* (DA) to evaluate the extent to which *confidence bias* was independently associated with these factors. In three of the cases (male sex, EASL and non-selective schooling) the relationship remained statistically significant ($p<0.01$) after adjusting for the DA score. In two cases the relationship became non-significant (non-White ethnicity and non-professional background).

In terms of the relationship with reported ethnicity; those reporting themselves as 'White' were least 'overconfident' and those reporting 'Black' ethnicity most 'overconfident' according to the *confidence bias* scores (Kruskall-Wallis test result; χ^2=267.74, p=0.0001). However, once overall performance on the DA subtest was controlled for this difference became non-significant (p=0.87). We also observed that *confidence bias* scores were moderately and inversely correlated with 'Extreme Response Style' (ERS- r=-0.46). This may have been because those more able candidates also rated their confidence as '5' and therefore more frequently observed making an 'extreme' response.

The results of the multilevel logistic regression indicated that those with higher *confidence bias* scores (i.e. indicating underconfidence) were much more likely to receive an offer from medical school (OR 11.97, 95% confidence intervals 9.55 to 15.01, p < 0.001). On univariable analysis, as previous research has indicated, a number of other background variables were significantly associated with the odds of receiving an offer for medical school. These included male sex, performance at decision analysis and the other subtests of the UKCAT, ethnicity, nonprofessional background, school-type attended and English language fluency. When the potential effects of these other background variables were controlled for it was interesting to note that the impact of the *confidence bias* score remained statistically significant (odds ratio 1.38, 95% confidence interval 1.04 to 1.82, p = 0.026). Thus, it appears that underconfidence in this context was a positive predictor of receiving an offer at medical school application.

The only statistically significant association between undergraduate performance and the *confidence bias* scores was that observed for the odds of passing the end of year one at first attempt. The odds ratio was 4.37 (95% confidence interval 1.54 to 12.42, p=0.006), indicating that the odds of passing first time, as opposed to another academic outcome, more than quadrupled for every standard deviation above the mean for applicants on the *confidence bias* score. That is, underconfidence, rather than overconfidence, was associated with an increased odds of passing year one. This effect appeared to be a linear, rather than a non-linear effect, as a quadratic term (*confidence bias* score squared) was not significant within a linear regression. Moreover, the effect appeared mainly independent of performance at decision analysis. When the effect of performance on this subtest of the UKCAT was controlled for the odds ratio remained significant (OR 3.24, 95% confidence intervals 1.08 to 9.72, p = 0.036). We observed no statistically significant relationship between the *confidence bias* score and the odds of passing year two at first attempt. Neither did we observe any significant association between the *confidence bias* scores and performance at either knowledge or skills-based assessments in year one year two of medical school.

In terms of a multivariable model, once the effects of one or more background variables independently associated with confidence bias (see above-language fluency, male sex and schooling) were controlled for the effect of *confidence bias* score itself was no longer statistically significant. Thus, it is likely that most, if not all of the association between confidence bias and the odds of passing year one are mediated by other, sociodemographic, variables, which themselves are associated with *confidence bias* scores.

As might be expected there was at least a modest correlation between *confidence bias* (the conventional way of modelling online confidence) and the novel *confidence judgment* score (r=0.24). However, as *confidence judgment* was calculated differently to *confidence bias* score there was a different relationship with both mean confidence rating and the SD of the ratings. In the case of *confidence judgment* the correlation with mean confidence ratings was r=-0.04 and with standard deviation r=0.23. This is in contrast to the relationship with *confidence bias* which correlated -.68 with the mean confidence ratings but only 0.11 with

SD. Thus *confidence judgement* appeared to have a weaker relationship with the mean self-ratings of confidence but a stronger relationship with the spread of these ratings. The weaker relationship with mean confidence is almost certainly due to controlling for the lift in ratings by the use of internal Z scores. However in theory this should also reduce the relationship with the spread of scores. However it is likely that the use of correlations in this case meant that candidates showing a wider spread of responses were also able to show larger correlation coefficients between their relative confidence about a particular item and their expected probabilities of a correct response. In contrast, candidates showing little variance in confidence ratings would not be able to achieve high correlation coefficients in this respect.

Interestingly, for the conventional method of calculating online confidence, there was a moderate to high correlation between the *confidence bias* score and the mean number of decision analysis items correctly answered (r=0.56). In contrast a more modest correlation between this index of overall performance on the DA subtest and *confidence judgement* was observed (r=0.29), suggesting it is less of pure proxy for overall ability at a cognitive task.

There was a trend of borderline statistical significance for the odds of being offered a place at medical school to be positively related to *confidence judgement* score. The odds ratio was 1.22 for this effect (95% confidence intervals 1.01 to 1.47, p=0.04). This could be interpreted as follows: for every standard deviation above the average for an applicant on *confidence judgement* the odds of being offered a place at medical school increased by around 20%. However, should be noted that on univariable analysis a whole range of other academic and sociodemographic factors related to the odds of an offer. Such factors included non-professional background, ethnicity, advanced school level qualifications etc. Subsequently when a multivariable model was built which controlled for the potential impact of these factors the predictive ability of *confidence judgement* was no longer statistically significant (OR 0.98, 0.80 to 1.20, p=0.8). Thus, it may be that *confidence judgement* serves as a proxy for other factors related to the odds of receiving an offer at application to medical school.

The relationship between performance at *knowledge* and *skills*-based exams in the undergraduate medical years and *confidence judgement* was even weaker than that observed for *confidence bias*. There were no statistically significant associations or even modest trends observed between *confidence judgement* scores and academic outcomes (including passing a year at first attempt).

Discussion

In these analyses we examined the relationship between online confidence and both academic outcomes and the probability of an offer of a place to study medicine. We are able to use an existing metric of ‘confidence bias’ utilised in the original pilot study by Pearson Vue. In addition we were able to explore the use of an experimental method to evaluate ‘confidence judgement’ using a relatively sophisticated approach to modelling self-appraisal of ability at a cognitive test. In line with previous research we found overconfidence was inversely related to cognitive performance. Our main finding was that *confidence bias* was modestly linked to subsequent academic performance, in the sense that those that were relatively underconfident were more likely to pass the first year of undergraduate medicine at first sitting. In addition underconfidence, as measured by the confidence bias score, was associated with an increased odds of receiving an offer to study medicine. Moreover this effect remained significant once the impact of their background demographic and academic variables were taken into consideration.

The novel measure of *confidence judgement* showed somewhat different properties to that of *confidence bias*. *Confidence judgement* showed a less strong relationship to cognitive ability. There was also a slightly different relationship with self-reported ethnicity. In contrast to *confidence bias*, there were no observed relationships between *confidence judgement* and academic performance. However *confidence judgement* was related to the odds of receiving an offer from medical school, though this effect was not independent of other educational and background factors.

It could be hypothesised that future doctors should ideally show neither overconfidence nor underconfidence. However, we did not observe any non-linear effects for confidence bias. This suggests that no such 'sweet spot' exists, at least relation to the measured outcomes examined in this study.

It is intriguing that 'underconfidence', as indexed by the *confidence bias* score, was associated with an increased probability of an offer from medical school. This effect seemed largely dependent of other educational, cognitive and background factors. It could be that candidates who appear overconfident are less likely to receive offers, as the vast majority of medical schools still use face-to-face interviews and/or group exercises as part of the selection process. Nevertheless, once in medical school this score had relatively little association with subsequent academic performance, and certainly no evidence of *independent* effects. Likewise the accuracy that one can appraise your own cognitive abilities, as indexed by the novel *confidence judgement* score, was associated with the odds of receiving offer from medical school, although unlike confidence bias, this effect did not seem to be independent of other factors.

In terms of selection policy, considering the measurement of either *confidence judgement* or *confidence bias*, there would seem to be a number of practical challenges with using such an approach in practice. Firstly it was noted in these data that a proportion of candidates did not vary in their responses in relation to their perceived confidence (i.e. there was zero variance). Thus such candidates provided no information about how they perceived their confidence related to their actual ability. It is difficult to see how such invalid response patterns could be discouraged or mitigated in practice. Indeed, if candidates believe that underconfidence was more desirable than overconfidence they could deliberately rate their confidence rather lower than they might otherwise. Thus it is difficult to defend against such 'social desirability' bias. A particular problem with using the conventional way of estimating confidence bias is that it seems to correlate relatively highly with actual cognitive ability (in this case r=0.51). Therefore there may be some doubts about the value that such measures would add, above and beyond cognitive ability metrics.

Thus, to conclude, in practice it may be more fruitful to concentrate on the attempted measurement of other aspects of personal qualities that may be deemed important to future clinical performance that are less strongly related to cognitive ability. In particular this facet of over or underconfidence might be better conceptualised as a component of interpersonal competency in this context. That is to say, individuals who come across as overconfident, arrogant or even narcissistic in interpersonal interactions are more likely to encounter difficulties in their future careers and workplaces than those who are more so socioemotionally competent. In this respect it may be worth revisiting some of the earlier work around emotional intelligence as an applied ability. In this sense emotional intelligence would be defined as being able to accurately identify emotional states in oneself and others, but also respond to them effectively. Such traits can be evaluated, to some extent, via situational judgment tests though more resource intensive approaches may be required to increase the precision by which such 'non-cognitive' attributes are measured.

References

McDaniel, M. A., J. Psotka, P. J. Legree, A. P. Yost and J. A. Weekley (2011). "Toward an understanding of situational judgment item validity and group differences." J Appl Psychol **96**(2): 327-336.
Pearson VUE (2017). Confidence Rating Data and Score. London, Pearson VUE.
Stankov, L., S. Morony and Y. P. Lee (2014). "Confidence: the best non-cognitive predictor of academic achievement?" Educational Psychology **34**(1): 9-28.