THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

A “Red Flag” system adds value to medical school admissions interviews

Citation for published version:
Ng, HY, Anderson, J, Marson, LP & Hope, D 2023, 'A “Red Flag” system adds value to medical school admissions interviews', *Medical Teacher*. https://doi.org/10.1080/0142159X.2023.2225724

Digital Object Identifier (DOI):
10.1080/0142159X.2023.2225724

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

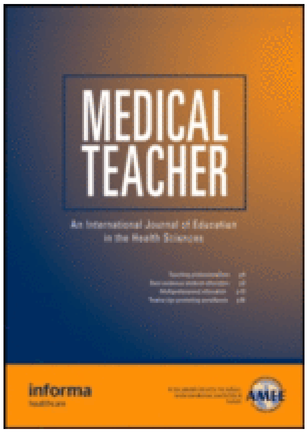
Published In:
Medical Teacher

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 04. Sep. 2023



A “Red Flag” system adds value to medical school admissions interviews

Journal:	*Medical Teacher*
Manuscript ID	CMTE-2022-0986.R1
Manuscript Categories:	Articles
Date Submitted by the Author:	n/a
Complete List of Authors:	Ng, Hak Yung; The University of Edinburgh, Medical Education Unit Anderson, Jane; The University of Edinburgh, Medical Education Unit Marson, Lorna; University of Edinburgh Division of Clinical and Surgical Sciences, Surgery Hope, David; The University of Edinburgh, Centre for Medical Education
Keywords (user):	Admissions, Interviews, Non-cognitive attributes
Keywords:	Selection < Management

SCHOLARONE™
Manuscripts

A "Red Flag" system adds value to medical school admissions interviews

Hak Yung Ng[1], Jane Anderson[1], Lorna Marson[2] and David Hope[1]

1. Medical Education Unit, The Chancellor's Building, The University of Edinburgh, College of Medicine and Veterinary Medicine, 49 Little France Crescent, Edinburgh, EH16 4SB, Scotland, United Kingdom
2. College of Medicine and Veterinary Medicine, Bioquarter, EH16 4SA, Scotland, United Kingdom

Address for Correspondence

Dr. David Hope

Medical Education Unit

The Chancellor's Building

The University of Edinburgh

College of Medicine and Veterinary Medicine

49 Little France Crescent, Edinburgh

EH16 4SB

Scotland, United Kingdom

david.hope@ed.ac.uk

For Peer-Review Only

Abstract

Introduction

Non-cognitive traits should be considered when selecting candidates to study medicine. However, evaluating these traits remains difficult. We explored whether measuring undesirable non-cognitive behaviour ("Red Flags") added value to a medical school admissions system. Red Flags included rudeness, ignoring the contributions of others, disrespectful behaviour, or poor communication.

Methods

Following an admissions interview testing non-cognitive attributes in 648 applicants to a UK medical school, we measured the association between interview score and Red Flag frequency. We tested linear and polynomial regression models to evaluate whether the association was linear or non-linear.

Results

In total, 1,126 Red Flags were observed. While Red Flags were concentrated among low-scorers, candidates in the highest- and second-highest deciles for interview score still received Red Flags (six and twenty-two, respectively). The polynomial regression model indicated candidates with higher scores received fewer Red Flags, but the association was not linear ($F(3,644) = 159.8$, $p = .001$, adjusted $R^2 = .42$).

Conclusions

The non-linear association between interview score and Red Flag frequency shows some candidates with desirable non-cognitive attributes will still display undesirable – or even exclusionary – non-cognitive attributes. Recording Red Flag behaviour reduces the likelihood such candidates will be offered a place at medical school.

For Peer Review Only

Keywords

Admissions, interviews, non-cognitive attributes, selection

Practice Points

We explored whether “Red Flags” (such as poor communication skills, rudeness, or disrespect to patients) should be included in medical school admissions systems

Even some academically excellent candidates with positive non-cognitive attributes displayed Red Flags

Without a Red Flag system, some of these candidates may have received offers as their undesirable behaviour would not have been recorded

For Peer-Review Only

Introduction

The recruitment of new doctors is an important topic for medical schools, students, healthcare workers, patients, and society (Quinlivan et al., 2010). If unsuitable applicants are selected, students may fail to graduate, experience disruption, or burnout. Meanwhile medical schools themselves may expend resources without providing new doctors to the workforce (Beer & Lawson, 2017).

Globally speaking, prior academic attainment, aptitude tests, reference letters, personal statements, various form of interviews, and even lottery systems have all been used in medical student selection (Prideaux et al., 2013). Substantial heterogeneity exists between and within different jurisdictions and countries (Wilkinson & Wilkinson, 2016). The optimal strategy for recruiting medical students remains unclear, and controversial (Poole et al., 2012), but developing fairer and more effective selection protocols is unarguably important (Swanwick, 2018). The selection decisions made today will shape the medical profession for fifty years to come (McGaghie, 2002), and effective selection systems are beneficial to society (Razack, 2016).

More than thirty years ago, the Edinburgh Declaration suggested that medical school should focus not only on academics, but on non-cognitive attributes (Stegers-Jager, 2018). In recent years, the healthcare field has explored how to evaluate non-cognitive attributes, beliefs, and values. This followed from the belief that it takes more than academic knowledge to make a competent doctor (Boulet & Durning, 2019; Harris & Owen, 2007; Lambe & Bristow, 2010). A wide range of attributes like integrity, creativity, maturity, effectiveness, critical thinking, professionalism, resilience, communication, teamwork, and empathy are recognized to be very important in daily clinical practice (Ayub, 2019; Patterson, 2018; Sebok & Syer, 2015). By clarifying to applicants which attributes are assessed on application, medical schools signpost

For Peer Review Only

what they really value (van Mook et al., 2009). More than eighty non-cognitive traits important to medical education have been documented in the literature **(Albanese et al., 2003; Sehiralti et al., 2010), and well-designed interview tools appear to predict performance on assessment within medical school (Barber et al., 2022).** However, it is very difficult and challenging to develop effective corresponding selection methods for these attributes (Kreiter, 2016; Patterson, 2018), and the validity and reliability of these non-cognitive assessment have been questioned (Goho & Blackman, 2006; Salvatori, 2001).

Besides this, not all non-cognitive traits are desirable. In addition to positive traits, researchers have sought to identify “Red Flags,” which make a candidate unsuitable to work as a doctor. In medical practice, Red Flags were originally used in the acute care of low back pain and appeared in the literature in the 1980s as a way to identify a major health problem. The underlying concept was then applied broadly across specialties (Ramanayake & Basnayake, 2018).

As applied to selection tools, Red Flags identify non-cognitive behaviours or attitudes which may be considered disqualifying even if a candidate otherwise scores highly. Lambe et al. (2018), exploring the inclusion of a Situational Judgement Test (SJT) in the dental student selection process, incorporated Red Flags into interviews. Applicants were red-flagged if they received a “no” from an interviewer on the question: “Would you like this applicant to be your dentist?” Interviewers could in this case, answer yes, maybe, or no. 29 applicants out of 189 were red-flagged in this way. However, they were unable to explore the reasons behind red-flagging, and applicants who received Red Flags appeared to be distributed across different SJT performance bands. This suggests that traditional interview scores, and disqualifying behaviour, are not necessarily highly correlated, and so Red Flags might measure a different aspect of applicant suitability.

For Peer Review Only

Sklar et al. (2015) compared a Multiple Mini Interview (MMI) system against a traditional interview system in a cohort of postgraduate applicants in Head and Neck Surgery. The assessors were given an opportunity to raise a Red Flag to express concerns about the suitability of applicants. However, there were no Red Flags raised for any of the 27 applicants. Bohrer-Clancy et al. (2018), explored the factors associated with negative outcomes of Emergency Medicine (EM) residency, and found Red Flags identified during EM clerkships predicted negative outcomes. Their definition of a Red Flag was a marked deficiency in the letter of recommendation, or written comments from attending physicians. However, among the 260 candidates analysed, there were only four red-flagged candidates out of 71 who had data available. The majority of candidate were not analysed, due to lack of data.

To summarise, there is evidence that well-designed interview systems can select candidates well- suited to studying medicine. Any improvements to interview systems which give assessors new insights into when and how to make offers will add value to the interview system itself. Alongside this, there is significant interest in the potential use of Red Flags to identify potentially disqualifying behaviour. However, it is not clear how common Red Flags are during selection processes, and it is not clear whether Red Flags add value beyond a traditional scoring system whereby applicants simply achieve higher scores for a better performance.

Our study explores whether a Red Flag system adds value by evaluating the association between interview score (where higher is better) and Red Flag frequency (where lower is better) and examines whether some high-scoring applicants still receive Red Flags. **If the system provides information that could not be obtained simply through a traditional scoring system, it will have added value to the interview process.**

For Peer Review Only

Methods

Research Approach and Study Design

Our framework for the study was derived from a pragmatic methodology in which we focused on the context of the problem rather than the specific method (Evans et al., 2011), and developed the study as part of a broader programme to investigate Widening Participation at our institution. The design of this study was selected to quantitatively measure the effects of recording what we defined as Red Flags, examples of which included an inability to communicate clearly, a tendency to ignore the contribution of other people, or being disrespectful towards patient groups. **All data within this study, including the interview scores, Red Flags, and survey results, were routinely collected for Quality Improvement purposes.**

Context

The undergraduate MBChB Medicine Programme at Edinburgh Medical School is a 6-year basic medical education program, designed to prepare students for work as a General Practitioner, hospital doctor, or academic. Students spend the first two years studying the fundamentals of medicine, health, ethics, and society, then undertake a year of research-based study in third year, before spending the remaining time on attachments developing the skills required to work as a new doctor. While the curriculum is developed locally, it is regulated by the General Medical Council (General Medical Council, 2018)

Participants and sampling

In this admission cycle (entry year 2020), there were 2,638 applicants to the MBChB Medicine Programme. 648 were invited and attended the interview. 32.6% were male while 67.4% were female. 497 applicants, 76.7% of applicants interviewed, were given offers. 237 applicants,

For Peer Review Only

36.6% of applicants interviewed, accepted the offer and were placed into the programme. Therefore, around one in every eleven applicants eventually studied on the programme. **All interviewed applicants in this cycle were included in the study.**

Procedure

Edinburgh Medical School remained the only medical school not conducting face to face interviews in the United Kingdom in 2017, largely for historical reasons. A review of the admissions process was undertaken in 2018, which included a review of medical school admissions processes across the country, observation of Multiple Mini-Interviews (MMIs) in another Scottish medical school and at the veterinary school in the same university, a review of the literature and personal correspondence. A survey of clinical teachers, including NHS and university employees was undertaken, followed by a workshop to which key contributors to the delivery of undergraduate medical education in Edinburgh were invited. Based on these discussions and results of the survey we introduced assessment days for students entering in 2020/21. Respondents to the survey and at the workshop were invited to prioritise core values and attributes as defined in the Medical Schools' Council guidance documents (Medical Schools Council, 2018).

For Peer Review Only

The three highest priorities were: motivation to study medicine and genuine interest in the medical profession, honesty, and the ability to treat people with respect. The lowest priority was academic ability, as this was assessed separately by academic grades.

A common thread throughout was the need to improve the diversity of our undergraduates, particularly socio-economic diversity, and so assessors carefully considered how students of some backgrounds might have fewer opportunities to engage in extra-curricular activities or work experience. The paucity of those wishing to become GPs was mentioned and a need to include more GPs in the selection process was highlighted.

The framework for assessment included three biographical interviews, aimed at gaining as much information about the candidate as possible, and focusing on one or two of the key attributes that were defined in the consultation process. The scoring system aligned to each of the attributes, and each candidate could score up to 7 for each of the assessed attributes. As a fourth station, applicants undertook a team activity, working in groups of three or four. **The first three stations were ten minutes long, whereas the fourth station was 30 minutes long. Descriptors were provided to assessors for each of the attributes, and they were asked to add free text to justify a Red Flag. As such, any significant failure on any of the attributes described earlier could constitute a Red Flag (a marker of concern), but no formal distinction was made between a "moderate" or "severe" Red Flag, and there was no defined list for assessors to use.** Additional information was provided in a free text box to support their use. The Red Flags were reviewed by the admissions team when making for offers. If a candidate received three or more Red Flags this gave cause for concern and the free text was carefully reviewed when considering whether to make an offer to the candidate. **No candidate with three or more red flags received an offer following interview in this dataset.**

The structure and timing were piloted with first year medical students. Selectors were volunteers who are clinically active as doctors or allied health professionals, involved in undergraduate teaching, and/or lay and patient representatives. All selectors attended a training event, which outlined the philosophy and process of the assessment days, **observed video exemplars for each station, and discussed how to use the Red Flags system for each**. For the first year, two selectors were present at each station, and marked independently. There was a briefing every morning, with a reminder about the structure and marking system, and members of the admissions team were available each day to respond to queries.

For Peer Review Only

Assessors were provided with a scoring scheme for each attribute up to a maximum of seven per attribute. Therefore, a station would provide a total of seven if assessing a single attribute, or fourteen if assessing two attributes. Each attribute received the same weighting. Red Flags were recorded as a simple numerical score. All stations were marked in the same way.

Statistical analyses

Statistical analyses were performed using R Statistical Software (Version 4.0.3; R Foundation for Statistical Computing, Vienna, Austria). For applicants interviewed, we collected data on interview scores and Red Flag frequency. All the data was completely deidentified and no personal data was used in the study. Ethical approval for the research was granted by the Medical Education Unit ethics committee.

There were two assessors for each station. We used the Intraclass correlation coefficient (ICC) for interrater reliability analysis: in a reliable selection test, different raters should broadly agree on the applicant's suitability, so reliability statistics should be at least moderate. The ICC is a statistical measure evaluating the level of agreement and correlation between measurements within the same class of data. It serves as a quantitative estimate of reliability, and the typical value falls between 0 and 1 (Liljequist et al., 2019). ICC estimates and their 95% confident intervals were calculated using R based on a mean-rating (k=2), absolute-agreement, 2-way random-effects model (Hallgren, 2012; Koo & Li, 2016; Shrout & Fleiss, 1979).

We then examined the association between interview score and Red Flag frequency by visual inspection and via polynomial regression. In a standard linear regression model, the association between predictor and outcome variable is assumed to take the form of a straight line. When

For Peer Review Only

using polynomial regression, it is possible to test the effects of curvilinear associations, for cases where the variables are associated but not in a form well-expressed by a straight line. In this model, we tested quadratic and cubic polynomials, after centring the station mark on the sample mean to reduce the likelihood of multicollinearity in the regression models. For a more detailed explanation of polynomial regression, see e.g. Bradley and Srivastava (1979).

This was important because if the association between interview score and Red Flag frequency was negative but not linear, it meant some applicants achieved high scores while still exhibiting potentially disqualifying behaviour. An a priori power calculation indicated the sample size was sufficient to detect small effects for all analyses (Cohen, 1992).

Since the focus of this study was solely on the association between interview score and Red Flag frequency, we do not discuss the relationship between interview scores and other components of admissions, or the threshold for receiving an offer.

Results

The descriptive statistics of the four stations can be found in Table 1. The mean score (and standard deviation – SD) for station one, two, three, and four were 6.07 (1.20), 5.98 (1.07), 5.05 (1.37), and 5.74 (1.20) respectively. The mean Red Flags for these four stations were less than 1. The SD of Red Flags ranged from 0.85-1.33. In total, 1,126 Red Flags were observed.

[Insert table 1 about here]

The Intraclass correlation coefficient (ICC) was used for the interrater reliability analysis for station one to four. ICC estimates and their 95% confident intervals were calculated using a mean-rating (k=2), absolute-agreement, 2-way random-effects model (Koo & Li, 2016; McGraw & Wong, 1996; Shrout & Fleiss, 1979). The ICC results of interview score are

For Peer Review Only

shown in Table 2. The reliability were Moderate, or Moderate to Good according to Koo and Li (2016). Using the proposed criteria by Cicchetti and Sparrow (1981), the reliability were Good, or Good to Excellent. Red Flag ICC result are shown in Table 3. The reliability were Moderate, or Moderate to Good according to Koo and Li (2016). Using the proposed criteria by Cicchetti and Sparrow (1981), the reliability were Fair to Good, or Good to Excellent. Therefore, both the interview score and the Red Flag frequency exhibited acceptable reliability.

[Insert Table 2 about here]

[Insert Table 3 about here]

The result showed a high level of skewness in which nearly half (48.5%) of candidates received no Red Flags at all. While 86.1% of candidates received fewer than five Red Flags, 1.7% of candidates received more than ten Red Flags. After dividing the applicants into deciles according to interview score, we noted that while a large number of Red Flags were given to the bottom decile (392), some were still given to both the highest decile (6) and the second-highest decile (22).

After this inspection, we visually examined the association of interview score against Red Flag frequency. We explored departures from linearity via a loess regression model (see Figure 1), in which the regression line was not forced into the form of a straight line. Generally, those with a higher interview score had fewer Red Flags – but the association did not seem linear.

[Insert Figure 1 about here]

To formally test for non-linear associations, we then ran a polynomial regression analysis, adding terms up to and including the 4th order as predictors. For these values, mean-centred

For Peer Review Only

scores were used as opposed to raw scores. Backwards elimination of the polynomial terms of interview score resulted in a model where interview score was expressed as a cubic function with coefficients: score, β = -.15, p = .001, score^2, β = 0.003, p = .001, and score^3 β = 0.0007, p = .001. The overall model fit statistics were F(3,644) = 159.8, p = .001, adjusted R^2 = .42 – a large effect size. The statistical analysis confirmed the visual inspection: the association between interview score and Red Flag frequency was not linear.

For Peer Review Only

Discussion

The interview scoring system was reliable, with examiners awarding consistent ratings to applicants across stations. Importantly, the Red Flags were also reliable; applicants who gained a Red Flag in one station were more likely to gain Red Flags in others. This supports the idea that the interview stations were measuring broad non-cognitive skills relevant to the selection criteria.

Low-scoring applicants received more Red Flags, but some very high-scoring applicants still exhibited concerning behaviour. The final, non-linear model supports this view: despite a large amount of variance shared between interview score and Red Flag frequency, they are not measuring identical constructs, and so incorporating Red Flags into an interview system is likely to add value when making selection decisions. Given the relatively small number of Red Flags observed in the top two deciles, highly competitive programmes may particularly benefit as they can then distinguish between two similarly high-scoring applicants when one is exhibiting disqualifying behaviour.

This study extends some of the previous research on non-cognitive selection methods. In line with some previous research (Lambe et al., 2018; Sklar et al., 2015), we were able to identify Red Flag behaviour. As with those studies, the average frequency of Red Flag behaviour was low, which confirms these are low-frequency, but potentially high-severity, events. More broadly, the fact even relatively high-scoring applicants sometimes exhibit concerning non-cognitive behaviour underlines the importance of non-cognitive evaluation in selection tools described by Stegers-Jager (2018). Our paper demonstrates that Red Flags can be used reliably to identify concerning behaviour. Red Flags can be incorporated into an interview system straightforwardly and add value to selection systems by explicitly considering negative, as well as positive, non-cognitive behaviour.

One notable finding was that our applicants had a relatively high number of Red Flags compared to other studies. Given the broad range of potentially concerning behaviour, and the fact that participants are sometimes applicants to medical school, sometimes medical students, and sometimes doctors, this may be expected. However, the "expected" frequency of Red Flags remains under-explored.

This study has a number of strengths. Firstly, the study was carried out on a brand-new interview system, and as such, there should have been no leakage of content and no practice effects from consulting with previous interviewees. The tool was reliable, and the sample size large enough to allow for the detection of small effects. The use of polynomial regression allowed for an effective test of non-linear associations, instead of relying on a simple linear model.

Despite this, there were limitations. This is a single-site study, in one jurisdiction. While the non-cognitive attributes considered desirable reflect those prioritised in other countries, some variance will be uniquely attributable to the local environment. Secondly, we were unable to compare different kinds of Red Flags, to see whether different forms of disqualifying behaviour were more prominent than others. **While assessors were trained, the possibility of implicit bias was not formally evaluated within this study.** Finally, this study did not

For Peer Review Only

link applicant score to future performance; the expectation that those with fewer Red Flags will make better medical students, and thereafter better doctors, is plausible but unconfirmed.

There are several logical options for future research. Firstly, a better exploration of the range and type of Red Flags may help selection committees understand the breadth of behaviour present in applicants and provide better guidance not just to interviewers, but to potential applicants as well. Testing whether some examiners give more Red Flags than others will help ensure reliability, while evaluating whether some applicant categories are given more Red Flags is important for understanding fairness in selection to medical school. Lastly, a longitudinal evaluation of what happens to applicants of different Red Flag frequencies and interview score will help selection committees better understand the validity of Red Flags as a concept. Collectively, these will enhance our understanding of selection methods and so, hopefully, improve the quality of the medical workforce.

In summary, the use of Red Flags added value to the interview system. Red Flags added unique information that could not be obtained through interview scores alone, providing new insights into applicants. This enhanced the decision-making processes of the admissions team.

For Peer Review Only

Acknowledgments

We gratefully acknowledge the assistance of the staff and students who helped develop the interview system.

Declaration of Interest Statement

No interest to declare.

References

Albanese, M. A., Snow, M. H., Skochelak, S. E., Huggett, K. N., & Farrell, P. M. (2003). Assessing personal qualities in medical school admissions. *Academic Medicine*, *78*(3), 313–321.

Ayub, R. (2019). Socially Accountable Medical Students: Selecting Medical Students for the 21st Century. *Journal of The Society of Obstetricians and Gynaecologists of Pakistan*, *9*(2), 70–71.

Barber, C., Burgess, R., Mountjoy, M., Whyte, R., Vanstone, M., & Grierson, L. (2022). Associations between admissions factors and the need for remediation. *Advances in Health Sciences Education*, *27*(2), 475–489.

Beer, C., & Lawson, C. (2017). The problem of student attrition in higher education: An alternative perspective. *Journal of Further and Higher Education*, *41*(6), 773–784.

Bohrer-Clancy, J., Lukowski, L., Turner, L., Staff, I., & London, S. (2018). Emergency medicine residency applicant characteristics associated with measured adverse outcomes during residency. *Western Journal of Emergency Medicine*, *19*(1), 106.

Boulet, J. R., & Durning, S. J. (2019). What we measure… and what we should measure in medical education. *Medical Education*, *53*(1), 86–94.

For Peer Review Only

Bradley, R. A., & Srivastava, S. S. (1979). Correlation in Polynomial Regression. *The American Statistician*, *33*(1), 11–14. https://doi.org/10.1080/00031305.1979.10482644

Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency*, *86*(2), 127–137.

Cohen, J. (1992). A Power Primer. *Quantitative Methods in Psychology*, *112*, 155–159.

Evans, B. C., Coon, D. W., & Ume, E. (2011). Use of Theoretical Frameworks as a Pragmatic Guide for Mixed Methods Studies: A Methodological Necessity? *Journal of Mixed Methods Research*, *5*(4), 276–292. https://doi.org/10.1177/1558689811412972

General Medical Council. (2018). *Outcomes for Graduates*. General Medical Council.

Goho, J., & Blackman, A. (2006). The effectiveness of academic admission interviews: An exploratory meta-analysis. *Medical Teacher*, *28*, 335–340.

Hallgren, K. A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23–34. https://doi.org/10.20982/tqmp.08.1.p023

Harris, S., & Owen, C. (2007). Discerning quality: Using the multiple mini-interview in student selection for the Australian National University Medical School. *Medical Education*, *41*(3), 234–241.

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

For Peer Review Only

Kreiter, C. D. (2016). A research agenda for establishing the validity of non-academic assessments of medical school applicants. *Advances in Health Sciences Education*, *21*(5), 1081–1085.

Lambe, P., & Bristow, D. (2010). What are the most important non-academic attributes of good doctors? A Delphi survey of clinicians. *Medical Teacher*, *32*(8), e347–e354.

Lambe, P., Kay, E., & Bristow, D. (2018). Exploring uses of the UK Clinical Aptitude Test-situational judgement test in a dental student selection process. *European Journal of Dental Education*, *22*(1), 23–29.

Liljequist, D., Elfving, B., & Roaldsen, K. S. (2019). Intraclass correlation – A discussion and demonstration of basic features. *PLOS ONE*, *14*(7), e0219854. https://doi.org/10.1371/journal.pone.0219854

McGaghie, W. C. (2002). Student selection. In *International handbook of research in medical education* (pp. 303–335). Springer.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

Medical Schools Council. (2018). *Statement on the core values and attributes needed to study medicine*. https://www.medschools.ac.uk/media/2542/statement-on-core-values-to-study-medicine.pdf

Patterson, F. (2018). Designing and Evaluating Selection and Recruitment in Healthcare. In *Selection and Recruitment in the Healthcare Professions* (pp. 1–26). Springer.

Poole, P., Shulruf, B., Harley, B., Monigatti, J., Barrow, M., Reid, P., & Bagg, W. (2012). Shedding light on the decision to retain an interview for medical student selection. *NZ Med J*, *125*(1361), 81–88.

For Peer Review Only

Prideaux, D., Roberts, C., Eva, K., Centeno, A., McCrorie, P., McManus, C., Patterson, F., Powis, D., Tekian, A., & Wilkinson, D. (2013). Assessment for selection for the health care professions and specialty training. In *International Best Practices for Evaluation in the Health Professions* (pp. 77–96). CRC Press.

Quinlivan, J. A., Lam, L. T., Wan, S. H., & Petersen, R. W. (2010). Selecting medical students for academic and attitudinal outcomes in a Catholic medical school. *Medical Journal of Australia*, *193*(6), 347–350.

Ramanayake, R., & Basnayake, B. (2018). Evaluation of red flags minimizes missing serious diseases in primary care. *Journal of Family Medicine and Primary Care*, *7*(2), 315.

Razack, S. (2016). 'Fairness' and student selection: The case for mandatory air quotes. *Medical Education*, *50*(6), 600–602. https://doi.org/10.1111/medu.12997

Salvatori, P. (2001). Reliability and validity of admissions tools used to select students for the health professions. *Advances in Health Sciences Education*, *6*(2), 159–175.

Sebok, S. S., & Syer, M. D. (2015). Seeing things differently or seeing different things? Exploring raters' associations of noncognitive attributes. *Academic Medicine*, *90*(11), S50–S55.

Sehiralti, M., Akpinar, A., & Ersoy, N. (2010). Attributes of a good physician: What are the opinions of first-year medical students? *Journal of Medical Ethics*, *36*(2), 121–125.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Sklar, M. C., Eskander, A., Dore, K., & Witterick, I. J. (2015). Comparing the traditional and Multiple Mini Interviews in the selection of post-graduate medical trainees. *Canadian Medical Education Journal*, *6*(2), e6.

Stegers-Jager, K. M. (2018). Lessons learned from 15 years of non-grades-based selection for medical school. *Medical Education*, *52*(1), 86–95.

For Peer Review Only

Swanwick, T. (2018). Understanding medical education. *Understanding Medical Education: Evidence, Theory, and Practice*, 1–6.

van Mook, W. N., van Luijk, S. J., O'Sullivan, H., Wass, V., Schuwirth, L. W., & van der Vleuten, C. P. (2009). General considerations regarding assessment of professional behaviour. *European Journal of Internal Medicine*, *20*(4), e90–e95.

Wilkinson, T. M., & Wilkinson, T. J. (2016). Selection into medical school: From tools to domains. *BMC Medical Education*, *16*(1), 1–6.

For Peer-Review Only

Table 1

Summary Statistics of the Stations

Station	Interview Score Mean (SD)	Range	RF Mean (SD)	RF Range
Station 1	6.07 (1.20)	3-14	0.39 (1.09)	0-8
Station 2	5.98 (1.07)	5-28	0.58 (1.33)	0-9
Station 3	5.05 (1.37)	3-28	0.55 (1.19)	0-7
Station 4	5.74 (1.20)	6-28	0.22 (0.85)	0-9

Note: RF = Red Flag. Raw interview scores were converted to a scale of 0-7

For Peer Review Only

Table 2

Interrater reliability analysis for Interview Score using Intraclass Correlation Coefficient (ICC)

	ICC2K (95% CI) **CI**: confidence interval	***Reliability (Koo and Li, 2016)***	***Reliability (Cicchetti and Sparrow, 1981)***
Station 1	0.72 (0.68-0.75)	Moderate to Good	Good to Excellent
Station 2	0.66 (0.61-0.70)	Moderate	Good
Station 3	0.73 (0.69-0.76)	Moderate to Good	Good to Excellent
Station 4	0.68 (0.64-0.72)	Moderate	Good

For Peer Review Only

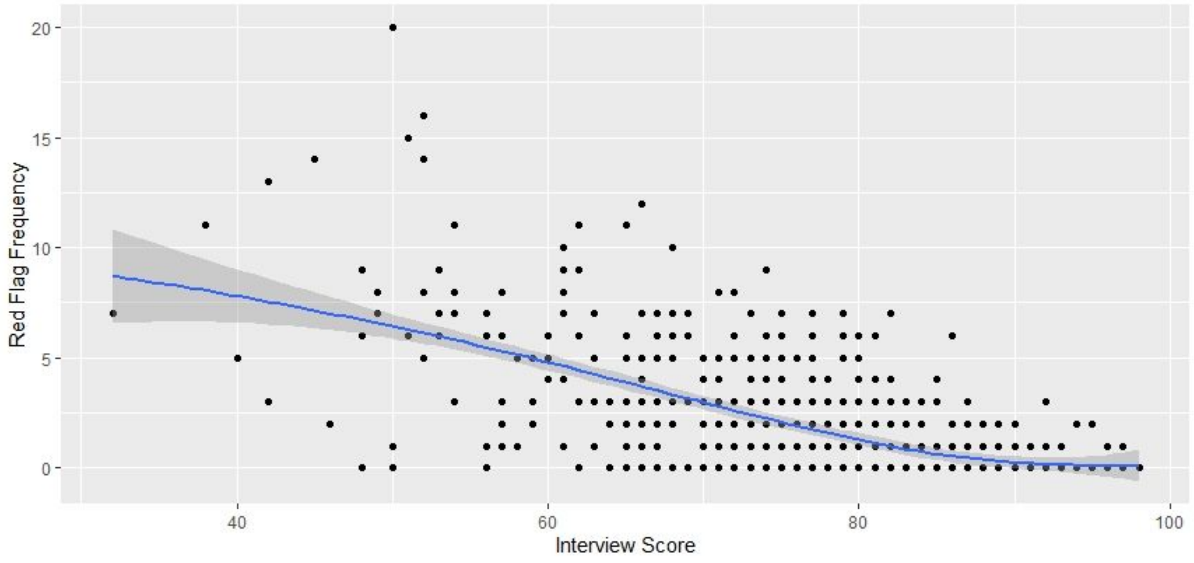
Table 3

Interrater reliability analysis for Red Flags using Intraclass Correlation Coefficient (ICC)

	ICC2K (95% CI) **CI**: confidence interval	***Reliability (Koo and Li, 2016)***	***Reliability (Cicchetti and Sparrow, 1981)***
Station 1	0.77 (0.74-0.80)	Moderate to Good	Good to Excellent
Station 2	0.73 (0.69-0.76)	Moderate to Good	Good to Excellent
Station 3	0.62 (0.57-0.67)	Moderate	Fair to Good
Station 4	0.74 (0.71-0.77)	Moderate to Good	Good to Excellent

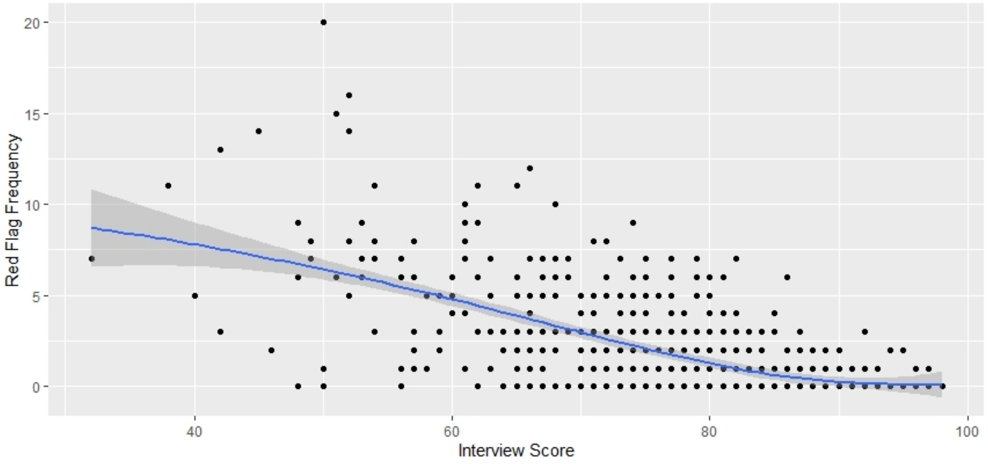
For Peer Review Only

Figure 1: Association between Interview Score and Red Flag frequency



Note: The blue line indicates the association between interview score and Red Flag frequency, and demonstrates the non-linearity of the association. **Data points with the exact same Red Flag Frequency/Interview Score are represented by a single dot.**

For Peer Review Only



Association between Interview Score and Red Flag frequency

227x107mm (96 x 96 DPI)

Reviewer: 1

Comments to the Author

1) Thank you for submitting this manuscript reporting your efforts to test a "Red Flag" system for medical school interviews. Early identification of candidates who demonstrate unprofessional or undesirable behaviors is of great interest to admissions teams and student affairs professionals. This work builds upon earlier work and extends our understanding of how the interview can contribute to the selection process.

> **Response:** We thank the reviewer for this comment. We have expanded the paper in a number of areas in response to reviewer comments.

2) Introduction: This is a succinct overview of the need for effective selection processes and the shift to consider non-cognitive factors in addition to tradition academic metrics. Red Flags are introduced as a practice that can be adopted for use in selection processes. The rationale for the use of Red Flags is clear and the problem and purpose statements are clear and focused. The constructs of non-cognitive attributes and judgement are introduced. Please articulate and describe the conceptual framework that guides this study.

> **Response:** We now note "Our framework for the study was derived from a pragmatic methodology in which we focused on the context of the problem rather than the specific method (Evans et al., 2011), and developed the study as part of a broader programme to investigate Widening Participation at our institution." We have made this as brief as possible but can expand if helpful.

3) The literature review includes many key papers and there is at least one current reference for each key topic (e.g., SJT, MMI). Please also see this recent paper for an excellent treatment of the relationship between selection procedures and professionalism.

Barber, C., Burgess, R., Mountjoy, M. et al. Associations between admissions factors and the need for remediation. Adv in Health Sci Educ 27, 475–489 (2022). https://doi.org/10.1007/s10459-022-10097-8

> **Response:** We thank the reviewer for this comment. We have included the suggested paper in the introduction.

4) Methods: Please start the section with a 1-2 sentence overview of the research approach and explain why this the appropriate design. This information is necessary before describing the context and participants. The information about the Quality Improvement process is interesting and relevant but explain whether/how this work launched the need for this study. Also, I was unclear about the source of the Red Flags until I reached the middle to end of the Procedures section. Please briefly explain this early on in the Methods section. Next, please add some information to explain what was done as a regular part of the selection process and what was done for this investigation. For example, the description of the survey and core values seemed disconnected. Please explain if the survey was done with intent to inform the study or was it something that had occurred as part of the Section Process QI effort?

Response: We have now modified the methods section as per comment 4 and 5. In particular, we note "All data within this study, including the interview scores, Red Flags, and survey results, were routinely collected for Quality Improvement purposes."

For Peer Review Only

5) I recommend following the traditional pattern, i.e., describe the research approach/paradigm, describe the research design, share the participants and context, describe the research methods and procedures or intervention, describe the measurement instrument, describe the participants and sampling procedures; discuss selection bias.... and then move to Data Analysis section.

> **Response:** We have substantially amended this section. We have added a new "research approach and study design" section outlining the framework, rationale for the study, and additional information on the source of Red Flags. We have separated the participants section into a "context" and "participants and sampling" section with additional information.

6) Data Analysis: I lack the expertise to evaluate all statistical aspects of this section, but the data analysis procedures are described with ample detail and seem appropriate to each analysis. The supporting statements about why specific tests were important are particularly helpful.

> **Response:** We thank the reviewer for this comment: we have aimed for concision throughout the analysis section.

7) Results: The results are communicated effectively and reinforce but not duplicate the Tables. Confirming the non-linearity of the relationship between interview score and Red Flag Frequency is an important result, as was establishing the reliability of the red flag assessment. I was astounded by the finding that more than half of the candidates received a red flag. What is known about the reviewers who assigned Red Flags? Are there hawks and doves? What is known about the recipients and the potential for implicit bias (e.g., if someone does not make consistent eye contact is that deemed disrespectful or poor communication?). The finding that 1.7% received more than ten flags is worrisome but not surprising.

> **Response:** We now note "One notable finding was that our applicants had a relatively high number of Red Flags compared to other studies. Given the broad range of potentially concerning behaviour, and the fact that participants are sometimes applicants to medical school, sometimes medical students, and sometimes doctors, this may be expected. However, the "expected" frequency of Red Flags remains under-explored." In limitations we now say "While assessors were trained, the possibility of implicit bias was not formally evaluated within this study."

8) Discussion: The Discussion effectively highlights the key points but does not overstate the findings. Limitations are noted, as are several appropriate next steps for this research. The practical significance of the findings (Red Flags are effective, can make the difference in top tier comparisons) are shared. Please add more information about the range of Red Flags observed. In the Procedures section there is only an abbreviated list (inability to communicate clearly, tendency to ignore..., being disrespectful). It will be helpful to be able to see the range of "severity" for the red flags. This is especially important as I anticipate some readers will ask if these were interview day faux pas due to interview anxiety -- or truly red flag behaviors that might (more work needed here though) predict professionalism problems upon matriculation.

> **Response:** In the methods we now note "Descriptors were provided to selectors for each of the attributes, and they were asked to add free text to justify a Red Flag. As such, any significant failure on any of the attributes described earlier could constitute a Red Flag (a marker of concern), but no formal distinction was made between a "moderate" or "severe" Red Flag, and there was no defined list for assessors to use."

For Peer Review Only

9) The tables and figures are easy to read and well-documented. The abstract is an accurate summary of the entire work.

> **Response:** We thank the reviewer for this comment.

Reviewer: 2

Comments to the Author

10) Ng & al have submitted a study based on a 4 station MMI, assessing the added value of a red flag mark in addition to the standard station mark.

Importantly, the research question posited in the title "A red flag system adds value to ....admissions interviews" has been neither addressed or proven. This would require a comparison between the results of the station scores plus/minus the red flag marks, to see how many cases the red flags changed the decision to offer or not. This does not occur in the paper, not least because the standard station marking is not described, and the red flags themselves are incompletely described.

> **Response:** We have now expanded our explanation of this, further explaining how we have utilised the term "value" and how we measured whether value was added. In the introduction, we now note:
>
> "To summarise, there is evidence that well-designed interview systems can select candidates well- suited to studying medicine. Any improvements to interview systems which give assessors new insights into when and how to make offers will add value to the interview system itself. Alongside this…"
>
> And later state:
>
> "If the system provides information that could not be obtained simply through a traditional scoring system, it will have added value to the interview process."
>
> So essentially, we argue the system adds value *because* it gives information otherwise unobtainable.
>
> Regarding the "changed offers," we have now expanded on this in response to comment 21.

11. How many types of red flags existed?

> **Response:** We have expanded on this as per comment 8: "Descriptors were provided to selectors for each of the attributes, and they were asked to add free text to justify a Red Flag. As such, any significant failure on any of the attributes described earlier could constitute a Red Flag (a marker of concern), but no formal distinction was made between a "moderate" or "severe" Red Flag, and there was no defined list for assessors to use."

12. Which types of red flags were awarded, and in which stations?

> **Response:** As outlined above in responses to 8 and 11, there was not a definitive list of Red Flags, and assessors used their judgment and training to decide whether a behaviour was a marker of concern.

13. How were selectors trained on what behaviours merited a red flag and what did not?

For Peer Review Only

Response: We now note that "All selectors attended a training event, which outlined the philosophy and process of the assessment days, observed video exemplars for each station, and discussed how to use the Red Flags system for each."

14. How were interview stations scored?

Response: We now note "Assessors were provided with a scoring scheme for each attribute up to a maximum of seven per attribute. Therefore, a station would provide a total of seven if assessing a single attribute, or fourteen if assessing two attributes. Each attribute received the same weighting. Red Flags were recorded as a simple numerical score."

15. How many selectors scored on each station to calculate an inter-rater correlation?

Response: This is noted in the procedure, but for clarity have added a statement to the statistical analyses as well: "There were two assessors for each station."

16 Apparently stations 1-3 were similar, focusing on biographical questions (p8l31) and station 4 was a group activity. Presumably the group activity was longer than each of the individual stations.

Response: We now note "The first three stations were ten minutes long, whereas the fourth station was 30 minutes long."

17. How was the group activity marked compared to the station 1-3 mark?

Response: We now note "All stations were marked in the same way."

18. Were the types or number of red flags different in station 4 compared to stations 1-3?

Response: No, the principles were the same for each station, for ease and clarity., and appeared to produce a comparable number of Red Flags.

19 The results of the MMI was that 77% of 648 interviewees were made offers

Response: Yes: this is a function of how UK medical school applications work, with a relatively large number of candidates declining to accept places at any given institution.

20 There appear to be much less than 648 data points in Fig 1 which shows the association between Red flags and interview score. Are there missing numbers?

Response: We now note "Data points with the exact same Red Flag Frequency/Interview Score are represented by a single dot." We could amend this by e.g. adding a jittered points function to the plot but we feel this would make the figure harder to interpret – but are happy to discuss if useful.

21. How many of the 149 interviewees who were NOT offered failed on >3 red flags?

Response: As a qualitative measure of concerning behaviour, there was no formal criteria for failing via Red Flags. Instead, these were discussed in greater detail by the applications team. The survey indicated that they found the additional information valuable, but, given the QI nature of the project and the first year of its operation, we did not automatically exclude any candidate based on Red Flags alone.

22. How many of the 497 interviewees who received offers were awarded >3 red flags, and what distinguished them from the group in the previous question 10?

For Peer Review Only

Response: We now note “No candidate with three or more red flags received an offer following interview in this dataset.” As noted above, this is not an automatic exclusion criteria and was considered alongside the interview scores, the range of assessors giving red flags, and academic results.

23. This study is remarkable for the 1126 Red Flags awarded to 648 interviewees, with over half of the interviewees receiving at least 1 Red Flag over 4 stations (p10l47 and p11l43-47). In this reviewer's experience, where red flags have been used to indicate behaviour meriting exclusion, there might only be 1-2 across an entire exam, even with higher numbers of students and more stations. Therefore, it appears that there is either a low threshold for awarding Red Flags, or that Red Flags can cover less concerning behaviours. This cannot be discerned without reference to Q1-3 above.

Response: We have now commented on this in the discussion: “One notable finding was that our applicants had a relatively high number of Red Flags compared to other studies. Given the broad range of potentially concerning behaviour, and the fact that participants are sometimes applicants to medical school, sometimes medical students, and sometimes doctors, this may be expected. However, the “expected” frequency of Red Flags remains under-explored.” We believe that the questions of “how many Red Flags should be given?” and “how variable should the amount of Red Flags given in different situations be?” are important future subjects for discussion.

24. It would be interesting if the authors concluded by answering the question posited in their title.

Response: We have expanded on this in the introduction. We note:

“To summarise, there is evidence that well-designed interview systems can select candidates well- suited to studying medicine. Any improvements to interview systems which give assessors new insights into when and how to make offers will add value to the interview system itself.”

And: “If the system provides information that could not be obtained simply through a traditional scoring system, it will have added value to the interview process.”

In the discussion we conclude: “In summary, the use of Red Flags added value to the interview system. Red Flags added unique information that could not be obtained through interview scores alone, providing new insights into applicants. This enhanced the decision-making processes of the admissions team.”

25. Exactly what value did the red flag system add to their process?

Response: We believe we have answered this in response to the reviewer comments above, especially in 23-24 – especially in terms of the additional information now available to assessors.

For Peer Review Only