Interviewer bias in medical student selection

Barbara N Griffin and Ian G Wilson

ABSTRACT

Objective: To investigate whether interviewer personality, sex or being of the same sex as the interviewee, and training account for variance between interviewers' ratings in a medical student selection interview.

Design, setting and participants: In 2006 and 2007, data were collected from cohorts of each year's interviewers (by survey) and interviewees (by interview) participating in a multiple mini-interview (MMI) process to select students for an undergraduate medical degree in Australia. MMI scores were analysed and, to account for the nested nature of the data, multilevel modelling was used.

Main outcome measures: Interviewer ratings; variance in interviewee scores.

Results: In 2006, 153 interviewers (94% response rate) and 268 interviewees (78%) participated in the study. In 2007, 139 interviewers (86%) and 238 interviewees (74%) participated. Interviewers with high levels of agreeableness gave higher interview ratings (correlation coefficient [r] = 0.26 in 2006; r = 0.24 in 2007) and, in 2007, those with high levels of neuroticism gave lower ratings (r = –0.25). In 2006 but not 2007, female interviewers gave higher overall ratings to male and female interviewees (t = 2.99, P = 0.003 in 2006; t = 2.16, P = 0.03 in 2007) but interviewer and interviewee being of the same sex did not affect ratings in either year. The amount of variance in interviewee scores attributable to differences between interviewers ranged from 3.1% to 24.8%, with the mean variance reducing after skills-based training (20.2% to 7.0%; t = 4.42, P = 0.004).

Conclusion: This study indicates that rating leniency is associated with personality and sex of interviewers, but the effect is small. Random allocation of interviewers, similar proportions of male and female interviewers across applicant interview groups, use of the MMI format, and skills-based interviewer training are all likely to reduce the effect of variance between interviewers.

MJA 2010; 193: 343–346

The use of interviews to select medical students is an internationally accepted practice because it enables assessment of important non-cognitive qualities, such as communication skills. Despite this popularity, there is evidence of considerable variance between individual interviewers' rating of interviewees,[1-3] and many Australian medical schools are adopting the multiple mini-interview (MMI) to reduce the effect of such variance. In the MMI, the applicant's score is an average of ratings of several mini-interviews, each of which is conducted by a different interviewer, "spreading" the effect of overly harsh or lenient individual interviewers. However, given the high-stakes nature of medical student selection, understanding the factors that might contribute to differences between interviewers is essential to reduce unwanted variance and thus improve the reliability and validity of ratings.

The main aim of this research was to investigate whether interviewer variance in the form of leniency bias (the tendency for some interviewers to rate interviewees more generously than most other interviewers) is related to the interviewer's personality or sex, including the interviewer being of the same sex as the interviewee. We also aimed to assess whether variance between interviewers is affected by the type of training they received.

Empirical tests of the widely accepted five-factor theory of personality[4] strongly support the claim that all facets of personality can be summarised by the so-called "big five" factors — extraversion, openness to experience, conscientiousness, neuroticism and agreeableness. We hypothesised that agreeableness would be the most likely of these factors to drive leniency bias because it describes the interpersonal qualities of generosity, sympathy, soft-heartedness and helpfulness.[4,5] The prosocial nature of highly agreeable people is likely to mean they take a lenient view of others.

Although research suggests that interviewer bias can be reduced by training,[6] the content and type of training reported in studies of medical selection interviews appear to vary considerably. We compare the variance in interviewer ratings after training that was predominantly knowledge-based with the variance in ratings after training that was predominantly skills-based.

METHODS

Participants

We analysed interview scores from two cohorts of MMI interviewees (one in 2006, the other in 2007) applying for admission to a new Australian medical school. Also in 2006 and 2007, we collected and analysed demographic and personality data from a cohort of each year's interviewers involved in the same MMI processes. The MMIs were run over 4.5 days each year, requiring 162 interviewers.

Measures

Interviewer rating

At each mini-interview (known as a "station"), interviewers made three ratings per interviewee using a 5-point Likert-type scale; each interviewer interviewed 20 applicants. A mean rating was calculated from the sum of the three ratings made by the interviewer for each of the 20 interviewees.

Interviewer personality

Interviewers completed the 20-item version[7] of the International Personality Item Pool,[8] measuring agreeableness, extraversion, neuroticism, conscientiousness and openness to experience. They were asked how accurately each item (eg, "sympathise with other's feelings") described them, using a scale from 1 for very inaccurate to 5 for very accurate.

Procedure

Applicants completed a 10-station MMI, which included one rest station. Each station lasted for 8 minutes and assessed a different quality. For example, Station 1 assessed applicants' motivation to study medicine and Station 9 assessed communication skills. Interview format also varied; some stations involved sets of questions about past behaviour and experience (behavioural interviews), others presented scenarios or film clips for comment, and at Station 9 applicants were required to explain something to a "patient" (role-played by an actor). There was one inter-

1 Hierarchical linear modelling showing the effect on multiple mini-interview (MMI) interviewees' scores of interviewer variance, sex of interviewee, sex of interviewer and interaction between sex of interviewee and sex of interviewer

MMI station*	Percentage of variance in interviewee score accounted for by between-interviewer variance		Sex of interviewee (beta main effect)		Sex of interviewer (beta main effect)		Interaction between sex of interviewer and sex of interviewee	
	2006	2007	2006	2007	2006	2007	2006	2007
1†	15.60%‡	3.15%	–0.59‡	ns	ns	—	ns	—
2	22.08%‡	10.77%‡	ns	ns	ns	ns	ns	ns
3†	20.59%‡	9.43%‡	ns	ns	ns	ns	ns	ns
4	8.50%‡	9.52%‡	ns	ns	ns	ns	ns	ns
5†	24.82%‡	3.18%	ns	–0.91§	ns	—	ns	—
6†	19.59%‡	12.27%‡	ns	–0.59§	ns	ns	ns	ns
7	6.17%‡	18.69%‡	ns	ns	–0.80§	1.52§	ns	ns
8	11.32%‡	19.55%‡	ns	ns	ns	ns	ns	ns
9	13.64%‡	3.11%	ns	ns	ns	—	ns	—

ns = not significant. * For security reasons, details of domains assessed at each station are not given; requests to authors for further information will be considered. † Station domain used in 2007 training. ‡ $P < 0.01$. § $P < 0.05$. (When interviewer variance is not significant, no interviewer factor is affecting interviewee score so no further analysis was conducted.) ◆

viewer per station. Ten applicants attended each MMI session and each interviewer worked for two sessions (ie, each interviewed 20 applicants).

All interviewers attended a 3-hour training session a month before the MMI. In 2006, the training was predominantly information-based, involving 2 hours of lecture about the rationale for including interviews in medical school student selection, information about the practical details of the MMI and how to score an applicant, the basics of behavioural interviewing, and instruction on avoiding bias. After a short break, the interviewers spent the remaining time in small groups practising using the rating scale and being given information about two MMI stations, with each small group studying different stations.

Feedback from interviewers indicated that they wanted more skills training. Therefore, the 2007 training sessions were restructured to be predominantly skills-based training. Interviewers practised rating "simulated" interviewees, comparing outcomes and discussing examples of good and bad responses, and they interviewed trainers and each other to learn to probe appropriately. Notably, this training used the actual content of four of the nine stations (Stations 1, 3, 5 and 6). In addition, interviewers attended a half-hour briefing immediately before interviewing at the 2007 MMI sessions, when they were given individual training on the content of the specific station they would be attending.

Analysis

It is essential to use multilevel modelling to account for the nested nature of the interview datasets on which studies such as ours are based.[9] When interviewees are rated by a subset of interviewers, they are "nested" under that subset. Analyses that disregard this multilevel component ignore dependencies between variables, artificially reduce standard errors and introduce correlated prediction errors. Not only does this violate statistical assumptions (eg, independence), but it increases the chance of finding significant results related to interviewer variables and decreases the chance of finding significant results related to individual (applicant) differences. Hierarchical linear modelling was therefore used (HLM 6.6 [SSI Scientific Software International, Lincolnwood, Ill, USA]), in addition to correlations and *t* tests for comparison of means. The threshold of significance was set at $P = 0.05$.

The research was approved by the institution's Human Research Ethics Committee.

RESULTS

In 2006, 153 interviewers (94% response rate) agreed to participate in the research and, in 2007, 139 (86%) participated (although of the latter, only 65% provided personality data). Interviewers were medical practitioners (18% in 2006, 14% in 2007); allied health workers (15% in 2006, 12% in 2007); university administrative personnel and lecturers from non-medical disciplines (39% in 2006, 35% in 2007); and local community members (27% in 2006, 40% in 2007). In 2006, 35% of the interviewers participating in the research were men and, in 2007, 33% were men.

We interviewed 342 applicants in 2006 and 321 in 2007; 268 (78%) of the former and 238 (74%) of the latter consented to participate in the research. The percentages of applicants who were men in 2006 and 2007 were 47% and 52%, respectively. The consent rates for interviewers and applicants combined were 86% in 2006 and 78% in 2007.

Effect of participants' sex

A mean score was calculated for each interviewee across the stations where he or she was interviewed by male interviewers and a second mean score was calculated for those stations where he or she was interviewed by female interviewers. Paired *t* tests were used to examine whether or not male or female interviewees received higher scores from male or female interviewers. In 2006, both male and female interviewees received higher scores from female interviewers than from male interviewers ($t = 2.99$, $P = 0.003$; $t = 2.16$, $P = 0.03$, respectively). In 2007, there were no significant differences between the average scores male or female interviewees received from male or female interviewers.

Multilevel analyses assessed the extent that the sex of interviewees contributed to the interviewee score at each station; whether the sex of interviewers contributed as a main effect to the interviewee score; and the interaction between sex of interviewer and sex of interviewee at each station (Box 1).

Female interviewees performed better than male interviewees at Station 1 in 2006 and at Stations 5 and 6 in 2007 (men and women did not differ in their total MMI score in either year[10]). Female interviewers differed from male interviewers only in the average score given to interviewees at Station 7. However, while women appeared to be more lenient at this station in 2006, they were less lenient than men in 2007. There was no significant interaction between sex of interviewer and sex of interviewee at any station in either 2006 or 2007, indicating

that neither female nor male interviewers were more lenient to interviewees of their own sex.

Effect of interviewer personality

Five-factor measurement of interviewer personality (agreeableness, extraversion, neuroticism, conscientiousness and openness to experience) yielded coefficient alphas of 0.58, 0.75, 0.61, 0.63 and 0.72, respectively, in 2006 and 0.71, 0.74, 0.70, 0.75 and 0.73 in 2007, with higher scores indicating higher levels of the five factors.

Correlations between interviewer ratings and personality are presented in Box 2. As hypothesised, we found agreeableness to be the only factor that significantly correlated with interviewer ratings in 2006. In 2007, interviewer neuroticism was also significantly correlated, with high neuroticism associated with lower (harsher) ratings. While the effect of interviewer personality was small,[11] accounting for less than 7% of the variance in scores, the strength of the correlations may have been due in part to the restricted range of the interviewer agreeableness scores (high with low variance).

Effect of training

A comparison of the effect of skills-based training with information-based training on the four stations that were the focus of the 2007 training (Stations 1, 3, 5 and 6) showed that the mean variance in interviewee scores attributable to interviewer differences was significantly reduced from 20.2% in 2006 to 7.0% in 2007 ($t = 4.42$, $P = 0.004$).

Overall interviewer effect

Multilevel analyses allowed us to assess the proportion of variance in interviewee scores accounted for by differences in interviewees (within-group variance) and differences in interviewers (between-group variance) at each station. The amount of variance in interviewee scores attributable to interviewers' differences ranged from 6.2% to 24.8% in 2006 and from 3.1% to 19.6% in 2007 (Box 1).

DISCUSSION

This study found that the personality and, to a lesser extent, the sex of interviewers are associated with the leniency of their ratings in a medical student selection MMI. Importantly, the results show that interviewers were not biased towards applicants of their own sex and there was evidence to suggest that type of training may reduce variance between interviewers.

Identifying stable individual characteristics that affect raters helps explain the observed "hawks-and-doves" pattern of rating, where "hawk" raters are thought to be more harsh in their rating style and "dove" raters more lenient. This pattern has been identified in both selection interviews and Objective Structured Clinical Examination (OSCE) assessment,[12,13] and found to be entrenched despite training of interviewers.[14] Given that personality traits are thought to be normally distributed, our finding that agreeableness in interviewers is associated with lenient interview ratings supports findings that the hawks-and-doves effect is normally distributed and stable over time[12] and suggests that personality testing could be used as a screening tool in high-stakes contexts for identifying those with the potential to be extreme raters. Unexpectedly, neurotic interviewers showed a tendency to rate more harshly in the 2007 interviews. In training that year, we had emphasised the problems of leniency, so perhaps in their anxiety to perform correctly, highly neurotic interviewers had over-compensated. Furthermore, the relationship of ratings with agreeableness could actually have been deflated because those scores were typically high with low variance. The interviewers in this study were all volunteers, and past research[15,16] has found that volunteers have higher levels of agreeableness. Agreeableness may therefore have a stronger effect in other rating situations, such as OSCE assessments, where raters are more likely to be recruited from among staff. Nevertheless, the effect of personality on interview scores was generally not substantial and only related to two of the "big five" factors; therefore, random allocation of interviewers will likely nullify most of the effect on applicants' scores. Given the indication that female assessors were somewhat more lenient, MMI panels should seek to have a similar proportion of men and women for each group of applicants.

In light of the debate about high levels of women entering medicine in Australia,[17] our results are important in showing that female performance at interview is not due to any bias from male or female interviewers.

The problem of rater leniency in medical selection interviews[18] was a factor leading to the development of the MMI.[3] By highlighting that significant variance in interview scores was accounted for by differences between interviewers, this study supports the use of the MMI format instead of panel or single interviews to mitigate against false-positive or false-negative decisions. Nevertheless, the amount of variance attributable to interviewers in our study was substantially less than that reported in studies of panel interviewers[1] and OSCE examiners,[19]

2 Mean scores for multiple mini-interview (MMI) interviewer ratings of interviewees and for interviewer personality traits, and relationships (correlation coefficients) between these values

	2006 mean score (SD)	2007 mean score (SD)	Correlation coefficient*					
			MMI score	Agreeableness	Extraversion	Neuroticism	Conscientiousness	Openness
MMI score	10.95 (1.29)	10.50 (1.19)	—	0.24‡	0.19	–0.25‡	–0.06	0.00
Agreeableness	4.23 (0.56)	4.11 (0.65)	0.26†	—	0.28†	–0.22‡	0.12	0.32†
Extraversion	3.24 (0.79)	3.29 (0.69)	–0.06	0.24†	—	–0.30†	–0.06	0.26†
Neuroticism	2.30 (0.66)	2.42 (0.71)	–0.08	–0.17‡	–0.16	—	–0.06	0.04
Conscientiousness	3.85 (0.68)	3.84 (0.72)	0.09	0.11	–0.08	–0.19‡	—	–0.02
Openness	3.30 (0.52)	3.18 (0.65)	–0.01	0.12	0.19‡	0.01	–0.11	—

* Correlation coefficients for 2006 data on lower diagonal (darker shading) and for 2007 data on upper diagonal (lighter shading). † $P < 0.005$. ‡ $P < 0.05$. ◆

and similar to or less than found in other MMI studies.[3,20] Furthermore, it appears that skills-based training of interviewers may reduce the variance between interviewers. Although these results need to be interpreted cautiously as we did not conduct a tightly controlled experiment and only present 2 years of data, they do challenge suggestions that training may be unnecessary.[20]

There is ongoing debate about the potential subjectivity of incorporating interviews into the medical student selection process. Our findings should alleviate some of that concern by showing that there is no evidence of sex bias and the effect of interviewer personality is relatively small. Further research is needed to investigate the effect of interviewer training, but we have provided initial evidence that skills training may increase the consensus between interviewers.

COMPETING INTERESTS

None identified.

AUTHOR DETAILS

Barbara N Griffin, BPsych(Hons), PhD, MAPS, Senior Lecturer[1]
Ian G Wilson, MB BS, PhD, FRACGP, Professor of Medical Education[2]

1 Psychology, Macquarie University, Sydney, NSW.
2 University of Western Sydney, Sydney, NSW.

Correspondence: barbara.griffin@mq.edu.au

REFERENCES

1 Harasym PH, Woloschuk W, Mandin H, Brundin-Mather R. Factors affecting the selection of students for medical school. *Acad Med* 1996; 71: S40-S42.
2 Roberts C, Walton M, Rothnie I, et al. Factors affecting the utility of the multiple mini interview in selecting candidates for graduate-entry medical school. *Med Educ* 2008; 42: 396-404.
3 Eva KW, Reiter HI, Rosenfeld J, Norman GR. The relationship between interviewers' characteristics and ratings assigned during a multiple mini-interview. *Acad Med* 2004; 79: 602-609.
4 Digman J. Personality structure: emergence of the five-factor model. *Annu Rev Psychol* 1990; 41: 417-440.
5 McCrae RR, Costa PT. A five-factor theory of personality. In: John LaPaop, editor. Handbook of personality: theory and research. 2nd ed. New York: Guilford Press, 1999: 139-153.
6 Pulakos ED, Schmitt N, Whitney D, Smith M. Individual differences in interviewer ratings: the impact of standardization, consensus discussion, and sampling error on the validity of a structured interview. *Personnel Psychol* 1996; 49: 85-102.
7 Donnellan MB, Oswald FL, Baird BM, Lucas RE. The mini-IPIP scales — tiny yet effective measures of the big five factors of personality. *Psychol Assess* 2006; 18: 192-203.
8 International Personality Item Pool: a scientific collaboratory for the development of advanced measures of personality and other individual differences. http://ipip.ori.org/ipip (accessed Aug 2010).
9 Sacco JM, Scheu CR, Ryan AM, Schmitt N. An investigation of race and sex similarity effects in interviews: a multilevel approach to relational demography. *J Appl Psychol* 2003; 88: 852-865.
10 Wilson I, Harding D, Yeoman N, et al. Lack of biases in the MMI. *Med Teach* 2009; 31: 959-960.
11 Cohen J, Cohen P, West SG, Aiken LS. Applied multiple regression/correlation analysis for the behavioral sciences. 3rd ed. Fort Worth: Harcourt Brace College Publishers, 2003.
12 McManus IC, Thompson M, Mollon J. Assessment of examiner leniency and stringency ('hawk–dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modeling. *BMC Med Educ* 2006; 6: 42.
13 Lawson DM. Applying generalizability theory to high-stakes objective structured clinical examinations in a naturalistic environment. *J Manipulative and Physiol Ther* 2006; 29: 463-467.
14 Harasym PH, Woloschuk W, Cunning L. Undesired variance due to examiner stringency/leniency effect in communication skill scores assessed in OSCEs. *Adv Health Sci Educ* 2008; 13: 617-632.
15 Graziano WG, Eisenberg NH. Agreeableness: a dimension of personality. In: R Hogan, Johnston J, Briggs S, editors. Handbook of personality psychology. San Diego, Calif: Academic Press, 1997: 795-824.
16 Gustavo C, Okun MA, Knight GP, de Guzman MRT. The interplay of traits and motives on volunteering: agreeableness, extraversion and prosocial value motivation. *Pers Individ Dif* 2005; 38: 1293-1305.
17 Laurence CO, Turnbull CO, Briggs NE, Robinson JS. Applicant characteristics and their influence on success: results from an analysis of applicants to the University of Adelaide Medical School, 2004–2007. *Med J Aust* 2010; 192: 212-216.
18 Mann WC. Interviewer scoring differences in student selection interviews. *Am J Occup Ther* 1979; 33: 235-339.
19 Iramaneerat C, Yudkowsky R. Rater errors in a clinical skills assessment of medical students. *Eval Health Prof* 2007; 30: 266-283.
20 Roberts C, Rothnie I, Zoanetti N, Crossley J. Should candidate scores be adjusted for interviewer stringency or leniency in the multiple mini-interview? *Med Educ* 2010; 44: 690-698.

(Received 14 Oct 2009, accepted 16 Jul 2010) ❑