Interviewer bias in medical student selection

Barbara N Griffin and Ian G Wilson

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.

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-
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.

### Effect of participants’ sex

A mean score was calculated for each interviewer 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 interviewer 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).

<table>
<thead>
<tr>
<th>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</th>
<th>Percentage of variance in interviewee score accounted for by between-interviewer variance</th>
<th>Sex of interviewer (beta main effect)</th>
<th>Sex of interviewee (beta main effect)</th>
<th>Interaction between sex of interviewer and sex of interviewee</th>
</tr>
</thead>
<tbody>
<tr>
<td>1†</td>
<td>15.60†</td>
<td>3.15%</td>
<td>-0.59†</td>
<td>ns</td>
</tr>
<tr>
<td>2</td>
<td>22.08†</td>
<td>10.77%</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>3†</td>
<td>20.59†</td>
<td>9.43%</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>4</td>
<td>8.50†</td>
<td>9.52%</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>5†</td>
<td>24.82†</td>
<td>3.18%</td>
<td>ns</td>
<td>-0.91§</td>
</tr>
<tr>
<td>6†</td>
<td>19.59†</td>
<td>12.27%</td>
<td>ns</td>
<td>-0.59†</td>
</tr>
<tr>
<td>7</td>
<td>6.17†</td>
<td>18.69%</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>8</td>
<td>11.32†</td>
<td>19.55%</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>9</td>
<td>13.64†</td>
<td>3.11%</td>
<td>ns</td>
<td>ns</td>
</tr>
</tbody>
</table>

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* < 0.05. (When interviewer variance is not significant, no interviewer factor is affecting interviewee score so no further analysis was conducted.)

### 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.
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, 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, and found to be entrenched despite training of interviewers. Given that personality traits are thought to be normally distributed, our finding that agreeableness in interviewers is more frequent than in the population is consistent with previous studies.

The problem of rater leniency in medical selection interviews was a factor leading to the development of the MMI. 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 and OSCE examiners.

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

<table>
<thead>
<tr>
<th>MMI score (SD)</th>
<th>2006 mean</th>
<th>2007 mean</th>
<th>Agreableness</th>
<th>Extraversion</th>
<th>Neuroticism</th>
<th>Conscientiousness</th>
<th>Openness</th>
</tr>
</thead>
<tbody>
<tr>
<td>MMI score</td>
<td>10.95 (1.29)</td>
<td>10.50 (1.19)</td>
<td>−</td>
<td>0.24†</td>
<td>0.19</td>
<td>−0.25‡</td>
<td>−0.06</td>
</tr>
<tr>
<td>Agreeableness</td>
<td>4.23 (0.56)</td>
<td>4.11 (0.65)</td>
<td>0.26‡</td>
<td>−</td>
<td>0.28†</td>
<td>−0.22‡</td>
<td>0.12</td>
</tr>
<tr>
<td>Extraversion</td>
<td>3.24 (0.79)</td>
<td>3.29 (0.69)</td>
<td>−0.06</td>
<td>0.24†</td>
<td>−</td>
<td>−0.30‡</td>
<td>−0.06</td>
</tr>
<tr>
<td>Neuroticism</td>
<td>2.30 (0.66)</td>
<td>2.42 (0.71)</td>
<td>−0.08</td>
<td>−0.17‡</td>
<td>−0.16</td>
<td>−</td>
<td>−0.06</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>3.85 (0.68)</td>
<td>3.84 (0.72)</td>
<td>0.09</td>
<td>0.11</td>
<td>−0.08</td>
<td>−0.19‡</td>
<td>−</td>
</tr>
<tr>
<td>Openness</td>
<td>3.30 (0.52)</td>
<td>3.18 (0.65)</td>
<td>−0.01</td>
<td>0.12</td>
<td>0.19†</td>
<td>0.01</td>
<td>−0.11</td>
</tr>
</tbody>
</table>

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

**Correlation coefficient**

- $P < 0.005$. † $P < 0.05$.
and similar to or less than found in other MMI studies. 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.

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
Ian G Wilson, MB BS, PhD, FRACGP, Professor of Medical Education

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

**Correspondence:** barbara.griffin@mq.edu.au

**REFERENCES**


(Received 14 Oct 2009, accepted 16 Jul 2010)