Evaluation of a resident selection process in plastic surgery

Mitchell H Brown BSC MD MEd FRCSC1, David Backstein MD MEd2, Glenn Regehr PhD3
1Division of Plastic Surgery, 2Division of Orthopedics, 3Department of Surgery and Centre for Research in Education, Faculty of Medicine, University of Toronto, Toronto, Ontario

Determined the attributes that make a successful resident is a fundamental task for postgraduate medical educators. Clarke and Wigton (1) have stated that the five most critical qualities to be examined in the selection of surgical residents are knowledge, judgement, technical skills, work habits and interpersonal skills. Greater insight regarding the qualities that are critical for resident success would allow the process of resident selection to be finely tuned, thereby minimizing speculation. As yet, however, these at-
tributes have not been clearly defined, and the process of resident selection continues to be an expensive and time-consuming combination of objective evaluation and ‘seat of the pants’ decision making.

The first step in resident selection is identifying candidates for personal interviews. Here, academic history seems to have a heavy weighting. Wagoner et al (2) conducted a survey of 210 residency program directors in Ohio to identify important issues in resident selection. Eighty-six per cent stated that preference is given to applicants who have excelled in an elective in the relevant specialty, and 46% select candidates based primarily on academic records. The perceived utility of the dean’s letter in resident selection was evaluated by Leiden and Miller (3) using a questionnaire sent to 124 medical schools. The authors found that little faith was put into the dean’s letter as a selection instrument due to a lack of standardization. Basic science grades were included less often than clinical marks, and the letters were often edited and summarized, reducing their credibility. Weaknesses were rarely mentioned and, leaves of absence or administrative actions were not always revealed.

Recognizing the importance of clinical performance data for selecting potential residents, DaRosa and Folse (4) evaluated a system designed to improve the validity and reliability of data derived from clinical clerkship performance. An objective, standardized assessment form was developed for the purpose of providing residency program directors with a method of comparing clinical clerks. The authors concluded that, while program directors appreciate this type of standardized form, clerkship directors are often unwilling to use it, especially for marginal students. Their hesitancy might be justified. Although the protocol was found to be a stable and reliable instrument, the form did not predict residency performance unless it was matched to very specific residency performance criteria. Therefore, more objective measures of academic achievement still appear to be the most popular method for initial candidate selection.

Once candidates have been selected for interviews, interviews appear to have a very important role in the final selections. Provan and Cuttress (5) surveyed 110 residency program directors in Ontario to determine their preferred methods of discriminating among applicants. The interview was most frequently identified as the primary discriminator; academic grades were ranked as second most important, with 38% preferring a numeric system, 28% a letter grade, 25% honours/pass/fail and 9% a pass/fail system. Consistent with these findings were the results of Baker et al (6) who analyzed the data from all applicants to the Department of Anaesthesiology at the Medical University of South Carolina over four years to compare application data of applicants who were accepted into the program with those of applicants who were rejected. While scores on standardized tests, academic grades and class ranks did not differ significantly between the two groups, the accepted candidates scored significantly higher in letters of reference and interviews, suggesting that these variables were critical in residency selection decisions.

The majority of residency programs use interviews in their admission process. Edwards et al (7) have stated that the interview serves four purposes: to gather information, to assist in decision making, to verify application data and to aid in recruitment. These authors stated that the interview’s most important function is to gather data that are not otherwise attainable through the standard written application. Foremost among these data is insight into the applicants’ interpersonal skills, leadership abilities and general level of motivation.

Despite the interview’s perceived ability to evaluate accurately qualities such as interpersonal skills, critics believe that it is less than valid due to a lack of standardization (8). Interviews can be structured, semistructured or unstructured. The structured format uses standardized questions for all applicants, provides sample answers for evaluation purposes and uses an interview panel. These standardization measures provide increased reliability and validity to the process (7). Interviewers should agree on the specifications being sought from the applicants, and the interview panel must consist of dedicated individuals who have worked together on multiple occasions (9). A larger panel allows deeper exploration of the candidate’s history and experience while decreasing the influence of any one interviewer, thus minimizing personal bias. In contrast, smaller panels allow greater cohesion among members and lend themselves to a more relaxed and informal atmosphere (9). Vickers and Reeve (9) recommended a well trained panel of two to three individuals as the ideal panel size.

Despite efforts by several authors to streamline the process of resident selection, many questions remain. The appropriate procedure for reviewing the applicant’s file and relative weighting of academic marks, references and dean’s letters is not yet clear. The reliability of the various components of the application selection process is unknown, and the effects of individual bias continue to be of concern. The relative importance of the interview component is often a contentious issue among program directors, as is the degree of structure and number of interviewers required.

By determining the reliability of application reviewers and interviewers, the number of consultants employed for these tasks could be minimized with a resultant savings of time and money. Finally, it would be of great interest and utility to examine how evaluators ‘gut’ ranking correlated with a systematic and objective approach to resident selection. This study examines an organized and thorough method of resident selection used by the Division of Plastic Surgery at the University of Toronto in 1997. By examining this methodology and its results, conclusions may be reached that will allow consolidation and improved efficiency of the residency selection process.

**PATIENTS AND METHODS**

The 11-person residency selection committee consisted of 10 faculty members and one resident representative from the division of plastic surgery. Included in the 10 faculty members were the chairperson and residency program director.

Fifty-one candidates completed the Canadian Resident Matching System (CARMS) application for plastic surgery
at the University of Toronto. The CARMS matching program provides standardized applications for each resident. These applications include candidate demographics, transcripts, letters of reference and a personal letter from the candidate. All 51 applications were reviewed and evaluated independently by both the chairperson of the division and by the program director. Each application was also evaluated independently by a third member of the residency selection committee (with each committee member rating five to six applications). Each of these three initial application scores was out of a total of 40 points, with 10 points being awarded for each of four categories: transcripts, research endeavours, references and overall quality of the application.

To select candidates for interviews, two sums were generated from the initial application scores. The first was the sum of the scores generated by the chairperson and program director (out of 80 points). The second was the sum of the scores generated by all three evaluators (the chairperson, the program director and the remaining committee member). This sum was out of 120 points. The top 10 candidates from each of the two lists were selected for interviews. Six candidates were in the top 10 of both lists and, therefore, the short list totalled fourteen candidates.

Each committee member was assigned to obtain more extensive information on one or two of the shortlisted candidates. This information included personal communications with references and program directors at the institutions where the candidates were attending medical school. Also obtained was information from faculty members with whom candidates had done electives as well as information from residents in the plastic surgery program who had worked with any of the candidates. The Residency Selection Committee met to share all information about the candidates to ensure that all committee members had equal information about each candidate. All committee members then took each of the 14 application packages and independently rescored them out of a total of 40 points using the same categories as before. These scores were identified as the ‘short list application scores’ for the candidates.

The interview took place approximately one month later. There were two interviews per candidate, with six committee members attending one interview and five committee members attending the second interview. Each interview was scheduled for 30 mins and had a semistructured format. Following each interview, evaluators were asked to provide an independent interview score for the candidate. Scores were not to be discussed or shared with other interviewers in the room. Each independent interview score was out of 40 points, with 10 points being awarded to each of four categories: general appearance and clarity of responses, enthusiasm, response content and overall presentation.

Further discussion was then undertaken by the entire Resident Selection Committee regarding the overall suitability of each candidate. Discussion points included the likelihood that the candidate would rank plastic surgery as their first choice, the potential for the candidate to undertake a successful career in research and education, and the location of the medical school that the candidate had trained at. A final ‘gut’ rank order from one to 14 was provided by each committee member. These rankings were summed, and the summed ranks, based on the final rank list, was formed.

**RESULTS**

**Reliability of the application scores**

An intraclass correlation coefficient (ICC) was calculated for the initial application scores by using the data from the chairperson and program director (who rated all 51 applications). The ICC for a single rater evaluating all applications was 0.64. The Spearman-Brown prophecy formula revealed that three raters evaluating all applications are required to obtain a reliability above 0.80. When rater was assumed to be a random factor (that is, assuming that a different rater is evaluating each application), the ICC was 0.58. Again, the Spearman-Brown prophecy formula revealed that three raters are required for each application to obtain a reliability above 0.80. This was confirmed by calculating the Cronbach’s alpha reliability of the three initial application scores. The reliability of the three scores was 0.80 despite the variety of raters who were contributing to the third score. Thus, it appears that it was not necessary to have the chairperson and program director rate all 51 applications. A similar reliability would have been obtained if the three evaluations for each of the 51 applications had been obtained by having the 11 members of the committee each evaluate approximately 14 applications.

The application scores were re-evaluated after additional information had been gained on a small subset of the candidates with a very restricted range of scores (only the top 14 applicants). Not surprisingly, therefore, the reliability of these ‘short list application scores’ were somewhat lower. By using the data from all 11 committee members, the ICC for a single rater evaluating all 14 applications was calculated to be 0.43. The Spearman-Brown prophecy formula revealed that six raters each rating all applications would be required to discriminate among these select applications with a reliability of at least 0.80 (Table 1).

**Reliability of interview scores**

To evaluate the reliability of the interview scores, an ICC was calculated for each interview (based on six raters in the first interview and five raters in the second interview) and for the scores collapsed across interviews (based on 11 raters). The reliability for a single rater in the first interview was 0.62, with three raters necessary to obtain a reliability of greater than 0.80. The reliability for a single rater in the second interview was 0.53, with four raters necessary to obtain a reliability of greater than 0.80 (Table 1). Similarly, the reliability of a single rater calculated across scores from both interviews was 0.55, with four raters necessary to obtain a reliability of greater than 0.80. That the reliability was not substantially lower when calculated across the two interviews suggests that the two interviews were quite consistent in how candidates were being rated. This was confirmed by the correlation between the summed interview scores from...
the two rooms, which was 0.77, and was estimated to be 0.88 when disattenuated for unreliability. In short, three to four raters evaluating a single interview seems to provide a fairly reliable measure of the candidates’ skills.

**Correlations with final ‘gut’ rankings**

The correlations between each stage of the application process and the final scores based on the committee members’ overall ‘gut’ ranking of the candidates can be seen in Table 2. Clearly, the initial application scores are not particularly useful in predicting final rankings. Again, this inability to predict the final outcome likely arises because the initial application scores are more directed at distinguishing individuals who should or should not receive an interview and are not aimed at distinguishing the best among those who deserve an interview.

By contrast, the short list application scores and the interview scores correlate quite well with the final ‘gut’ ranks, suggesting that the committee members are making use of this information when making their final decisions. Further, a multiple regression of short list application scores and interview scores predicting final ‘gut’ rank indicated that a score weighted 61.5% application and 38.5% interview was the most effective predictor of final ‘gut’ rank, with a multiple correlation of 0.96.

**DISCUSSION**

The evaluation of candidates for resident positions continues to be a very difficult challenge for training programs in all disciplines. The present study analyzed many of the basic psychometric principles as they apply to the selection process in an attempt to improve our ability to discriminate reliably among various applicants.

In the CARMS, there can be a large number of applicants and form an opinion once all of the information had been obtained. The committee thought that this degree of standardization with respect to the evaluation of applicants provided a level of comfort for such an important decision-making process. A large amount of information was collected on each applicant, including application, interview, personal commun-

<table>
<thead>
<tr>
<th>Number of raters</th>
<th>Interview Fixed</th>
<th>Initial application Fixed</th>
<th>Initial application Random</th>
<th>Short list application Fixed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.55</td>
<td>0.64</td>
<td>0.58</td>
<td>0.43</td>
</tr>
<tr>
<td>2</td>
<td>0.71</td>
<td>0.78</td>
<td>0.73</td>
<td>0.60</td>
</tr>
<tr>
<td>3</td>
<td>0.79</td>
<td>0.84</td>
<td>0.81</td>
<td>0.69</td>
</tr>
<tr>
<td>4</td>
<td>0.83</td>
<td>0.88</td>
<td>0.85</td>
<td>0.75</td>
</tr>
<tr>
<td>5</td>
<td>0.86</td>
<td>0.90</td>
<td>0.87</td>
<td>0.79</td>
</tr>
<tr>
<td>6</td>
<td>0.88</td>
<td>0.91</td>
<td>0.89</td>
<td>0.82</td>
</tr>
<tr>
<td>7</td>
<td>0.90</td>
<td>0.93</td>
<td>0.91</td>
<td>0.84</td>
</tr>
<tr>
<td>8</td>
<td>0.91</td>
<td>0.93</td>
<td>0.92</td>
<td>0.86</td>
</tr>
<tr>
<td>9</td>
<td>0.92</td>
<td>0.94</td>
<td>0.93</td>
<td>0.87</td>
</tr>
<tr>
<td>10</td>
<td>0.92</td>
<td>0.95</td>
<td>0.93</td>
<td>0.88</td>
</tr>
</tbody>
</table>

*Fixed Reliability assuming that ‘rater’ is a fixed factor (same rater evaluating all candidates). Random Reliability assuming that ‘rater’ is a random factor (different raters evaluating each candidate).*

...equipped with a set of standards and guidelines could be used to rate a smaller subset of applications. Once a short list of applicants has been selected, a larger number of raters is required to maintain a reliability of greater than 0.80. In this study, six raters provided that level of reliability. The difference is most likely selection of the top subset of applicants resulting in a more homogeneous group, which is known to affect reliability.

The interview process is costly in time and in human resources, and puts a great deal of stress on the applicants. There are, however, a great number of advantages to the interview process, including the ability to meet an individual face to face, to discuss issues that are not commonly identified on a written application and to evaluate an individual’s interpersonal skills. In this study, two separate 30 min interviews were undertaken for each applicant. The high correlation between the two interviews suggests that only one interview was required. As well, only four raters were required to obtain a reliability of greater than 0.80.

Finally, the ‘gut’ ranking determined the committee’s selections. The correlation of this ‘gut’ rank to both the application scores and the interview scores was extremely high, indicating consistency across all levels of evaluation. It is also interesting to note that a weighting of approximately 60% for the application and 40% for the interview most closely predicted the eventual final ‘gut’ rank. Therefore, it appears that, at least for this group, the interview process did not unduly influence the final decisions. Other data, such as academic achievement, had at least as large a role in the final rankings generated by the individual committee members.

In general, members of the selection committee found this process to be quite workable. It allowed the members to focus clearly on the applications and interviews independently and form an opinion once all of the information had been obtained. The committee thought that this degree of standardization with respect to the evaluation of applications provided a level of comfort for such an important decision-making process.
communication with references and personal experience with residents. The correlation of our final rank order with our initial ranking at the beginning of the selection process was 0.38. It is clear that the process outlined by the selection committee had a profound effect on our eventual rankings.

In general, each of the steps taken during the selection process described here seems to have been an important, useful and legitimate step. Each step generated reliable measures of the residents, and each step was used to a reasonable and appropriate degree in the final decision process by the committee members.

CONCLUSIONS
The present study analyzed the resident selection process in the Division of Plastic Surgery at the University of Toronto. Applications can be screened initially with high reliability using three evaluations per applicant provided by various selection committee members. Further review of the applications for the short list candidates required six individuals to obtain a reliability of greater than 0.80. Only one interview was required per candidate, and four evaluators were required to obtain a reliability of greater than 0.80. Finally, if application and interview scores are to be summed to obtain a final score, a weighting of 60% for the application and 40% for the interview most closely predicts the final assessment.

ACKNOWLEDGEMENTS: The authors thank the residency selection committee in the Division of Plastic Surgery, University of Toronto for their time, patience and enthusiastic participation in this study.

REFERENCES

TABLE 2
Correlations between various stages of resident screening process

<table>
<thead>
<tr>
<th></th>
<th>Initial application</th>
<th>Short list application</th>
<th>Interview</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial</td>
<td>–</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>application</td>
<td>score</td>
<td>score</td>
<td></td>
</tr>
<tr>
<td>Short list</td>
<td>0.63</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>application</td>
<td>score</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interview</td>
<td>0.19</td>
<td>0.75</td>
<td>–</td>
</tr>
<tr>
<td>score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Final ‘gut’</td>
<td>0.37</td>
<td>0.91</td>
<td>0.85</td>
</tr>
<tr>
<td>ranks</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>