

## Evaluating surgical resident selection procedures

Michael K. Gilbert, M.D., M.Ed.<sup>a,\*</sup>, Michael D. Cusimano, M.D., M.H.P., Ph.D.<sup>b</sup>,  
Glenn Regehr, Ph.D.<sup>c</sup>

<sup>a</sup>Department of Orthopaedic Surgery, University of Toronto, Toronto, ON, Canada

<sup>b</sup>Department of Neurosurgery, University of Toronto, Toronto, ON, Canada

<sup>c</sup>Centre for Research in Education, University of Toronto, Toronto, ON, Canada

Manuscript received October 19, 2000; revised manuscript November 22, 2000

---

### Abstract

**Purpose:** The purposes of this study were to develop and assess a rating form for selection of surgical residents, determine the criteria most important in selection, determine the reliability of the assessment form and process both within and across sites, and document differences in procedure and structure of resident selection processes across Canada.

**Methods:** Twelve of 13 English-speaking orthopedic surgery training programs in Canada participated during the 1999 selection year. The critical incident technique was utilized to determine the criteria most important in selection. From these criteria a 10-item rating form was developed with each item on a 5-point scale. Sixty-six candidates were invited for interviews across the country. Each interviewer completed one assessment form for each candidate, and independently ranked all candidates at the conclusion of all interviews. Consensus final rank orders were then created for each residency program. Across all programs, pairwise program-by-program correlations for each assessment parameter were made.

**Results:** The internal consistency of assessment form ratings for each interviewer was moderately high (mean Cronbach's alpha = 0.71). A correlation between each item and the final rank order for each program revealed that the items work ethic, interpersonal qualities, orthopedic experience, and enthusiasm correlated most highly with final candidate rank orders ( $r = 0.5, 0.48, 0.48, 0.45$ , respectively). The interrater reliabilities (within panels) and interpanel reliabilities (within programs) for the rank orders were 0.67 and 0.63, respectively. Using the Spearman-Brown prophecy formula, it was found that two panels with two interviewers on each panel are required to obtain a stable measure of a given candidate (reliabilities of 0.80). The average pairwise program-by-program correlations were low for the final candidate rank orders (0.14).

**Conclusions:** A method was introduced to develop a standard, reliable candidate assessment form to evaluate residency selection procedures. The assessment form ratings were found to be consistent within interviewers. Candidate assessments within programs (both between interviewers and between panels) were moderately reliable suggesting agreement within programs regarding the relative quality of candidates, but there was very little agreement across programs. © 2001 Excerpta Medica, Inc. All rights reserved.

*Keywords:* Medical education; Surgical residency; Selection; Critical incident technique

---

Despite its importance in shaping the careers of individuals and the health care of the future, the process of residency selection remains an area steeped in tradition and sparse in scientific rigor [1]. There has been some research performed on the resident selection process. Wagoner et al [2], for example, conducted a survey of residency program directors in order to determine the issues most important in selecting residents for interviews. Eighty-six percent stated that they would give preference to students who had done well in an elective in that specialty, while 46% stated that they selected

candidates for interviews based primarily on their academic records. Clarke and Wigton [3] attempted to develop an objective rating system for selecting surgical residents. They found that the categories of knowledge and judgment accounted for 50% of the weighting for resident selection. Tarico et al [4], using a procedure they called the critical incident technique for radiology, defined what was required for successful resident performance, and then designed a selection form based upon evaluating these characteristics. They were able to conclude that the scores on interviews using techniques based upon critical behavioral descriptors for candidates correlated well with residency performance.

---

\* Corresponding author. Tel.: (416) 340-4777; fax: (416) 340-3792.

Despite these occasional efforts, however, there is still relatively little known about the reliability of various components of the resident selection process and there is little consensus regarding the appropriate weighting of these components for the process of resident selection. Determining these parameters and using them to develop a standardized residency application rating form has the significant potential of contributing to the standardization and improvement of the selection of residents, and allows for comparisons between programs.

The purposes of this study were to develop and assess a rating form for selection of surgical residents (grounded in criteria for successful performance), determine the criteria most important in selection, determine the reliability of the assessment form and process both within and across sites, and document differences in procedure and structure of resident selection processes across Canada. The specific information gained from this particular study is likely to inform the orthopedic community regarding the important parameters of resident selection in their domain. At least as important, however, we believe that we are providing a prototype methodology for establishing these parameters in other surgical and other medical domains.

## Methods

### *Development of the measure*

The critical incident technique as described by Tarico et al [4], was utilized at the beginning of the study period in order to define the characteristics considered essential for successful selection into an orthopedic surgery residency training program. This involved the administration of mailed questionnaires to program directors and selection committee members across the country. Each participant was asked to identify the current selection practices in use at each institution, and to list (in order of importance) the characteristics considered essential for successful selection into, and completion of, an orthopedic surgery residency training program.

The responses were collated and characterized into groups, generating a consensus list of essential characteristics. These characteristics were redistributed to participants to assess whether the essence of important characteristics and categories had been captured. Final changes were then made to the list of characteristics, which was used to generate a concise rating form to be used for candidate selection on the day of interviews by all participating programs. Final consensus agreement was obtained from program directors and selection committee members on the content and structure of the rating form. A 10-item rating form was produced that included 9 specific dimensions of performance and a final rating of overall performance. Each item was rated on a 5-point scale with 1 representing poor performance on a given item and 5 representing outstanding performance on a

Table 1  
Selection assessment form items and their average correlation with the candidate's ranking

Item	Correlation with candidate ranking
1. Academic record/intelligence (grades, awards, previous degrees)	0.37
2. Curiosity (research experience, publications, sense of enquiry)	0.37
3. Orthopedic experience (electives, reference letters, familiarity with orthopedic lifestyle)	0.45
4. Enthusiasm (desire for orthopedics at institution)	0.50
5. Work ethic (self-motivated, hard-working, goal oriented)	0.48
6. Extracurricular activities/outside interests (roundedness of the individual)	0.34
7. Interpersonal qualities (communication skills, ability to work as a team member)	0.48
8. Ethical character (integrity, honesty, forthright, caring)	0.40
9. Psychomotor skills/manual dexterity	0.40
Sum of 9 items	0.69
10. Overall impression	0.76

given item (see resident selection assessment form headings, Table 1). The rating form was used during the interviewing process in the part of the study described below.

### *Assessment of the selection process*

On the day of interviews each program proceeded with its normal structure of interviews and resident selection. In addition, however, each interviewer was asked to independently complete the rating form for each candidate assessed immediately after the interview. At the conclusion of all interviews, each interviewer was asked to independently rank all candidates assessed that day (called "interviewer rank order"), prior to discussion with other members of the selection committee. After discussion among all interviewers a "consensus final rank order" was created and recorded by each program.

### *Statistical analysis*

Statistical analyses were divided into four sections.

#### *Internal consistency of the form*

The first set of analyses assessed the psychometric properties of the form itself, measuring internal consistency of the 9 dimensions of competence that were being assessed on the form, and the contribution of each dimension to the overall ranking of the candidate. To do this a Cronbach's alpha coefficient was calculated across the 9 dimensions for each independent rater. The average alpha across raters was calculated. Similarly, a correlation between each of the 9

dimensions (and the overall score) and the final rankings generated by the individual rater were computed, and the average correlation across raters was calculated for each dimension.

#### *Within panel interrater reliability*

Within each program interviewers were divided into various numbers of interview panels with several raters on each panel. Within each panel (for all but one site) the interviewers on the panel interacted with the candidate together, but then independently assessed the candidate using the form. Thus, for each panel it was possible to calculate an interrater reliability coefficient to determine the level of agreement across raters within a panel viewing the same set of performances. The average interrater reliability was calculated across panels. However, different panels included different numbers of interviewers, so in order to average the numbers sensibly, it was necessary to use a reliability measure that was insensitive to the number of raters. Thus, the reliability measure used to assess interrater reliability within a panel was the single-rater intraclass correlation coefficient (ICC), which is the equivalent of a Cronbach's alpha for a single rater. Once the average ICC was calculated, the Spearman-Brown prophecy formula was used to determine the number of independent raters on a panel necessary to achieve an alpha coefficient of 0.80.

#### *Within program interpanel reliability*

For programs where the same set of candidates were evaluated by more than one independent panel, it was possible to calculate the reliability of the scores generated by the panels. As for the within panel interrater reliability, an interpanel ICC was calculated for each program where there was more than one panel interviewing the same set of candidates. The average ICC was calculated across programs, and the Spearman-Brown prophecy formula was used to calculate the number of independent panels necessary to obtain an interpanel alpha of at least 0.80.

#### *Between program reliability*

A relatively consistent group of candidates were interviewed at most or all programs (ie, those with an interest in the specialty of orthopedics). Thus, for most pairs of programs, there were a reasonable number of candidates that were interviewed at each program. In order to calculate the reliability of the selection process between orthopedic programs, pairwise program-by-program correlations were made between rating form item results and consensus final rank order results for each candidate across the country, and the average correlation across all program pairs was calculated.

## **Results**

A total of 66 candidates from the Canadian Resident Matching Service (CaRMS) were invited for interviews

across the country. Twelve of 13 English-speaking orthopedic training programs in Canada participated during the 1999 selection year. The interview structures and contents differed among the 12 programs. In 8 programs each candidate was seen and interviewed by every member of the selection committee. In 4 programs each candidate was not seen and interviewed by every member of the selection committee, owing to the fact that candidates were not assigned to interview with each interview panel. The content of the interviews differed across programs. Four programs utilized a structured interview format, whereas 8 programs utilized an unstructured interview format. There was considerable variability in the number of interviewers used per program (mean 9.0, SD 2.9, range 4 to 16), and the number of panels of interviewers used per program (mean 3.7, SD 2.3, range 1 to 8). The mean duration of interviews was 22 minutes (SD 5.4 minutes, range 15 to 30).

#### *Internal consistency of the rating form*

As a measure of the internal consistency of the 9 items on the form, the average Cronbach alpha calculated across all interviewers was a moderately high 0.71. The average correlations between each of the 9 items and the consensus final rank order of candidates are presented in Table 1. The items enthusiasm, work ethic, interpersonal qualities, and orthopedic experience correlated most highly with the consensus final rank orders of candidates, but all items correlated moderately well. Given the reasonable internal consistency of the items on the form and the fact that there was not a substantial difference in the importance of any one item in generating the overall rank, it was decided that summing the 9 items to obtain a "total" score on the form was reasonable. The correlation of this sum with the final rank was 0.69.

#### *Interrater reliability within panels*

The interrater reliabilities (within panels, across raters) were calculated in order to determine the extent to which interviewers agreed on the interview performances they witnessed. The mean single-rater ICC for the total score, the overall impression, and the candidate ranking can be seen in Table 2. Based on these numbers, it appears that as the assessment becomes more global (from items to overall to ranks), candidates are differentiated more widely, and thus the measures become more reliable. Using the Spearman-Brown prophecy formula, it was determined that the average of two independent raters observing a single interview would be sufficient to obtain an interrater reliability of 0.80 for the ranking of candidates.

#### *Interpanel reliability within programs*

The mean intraclass correlation coefficients (ICCs within programs, across panels) were calculated in order to deter-

Table 2  
Average reliabilities of the ratings generated using the form

Measure	Internal consistency (Cronbach's alpha)	Interrater reliability within panel (ICC)	Interpanel reliability within program (ICC)	Interprogram reliability (mean pairwise correlation)
Sum of 9 items	0.71	0.45	0.58	0.18
Overall impression scores	—	0.49	0.63	0.18
Individual interviewer rank order	—	0.67	0.63	0.16
Final consensus rank order	—	—	—	0.14

ICC = intraclass correlation coefficient.

mine the degree of consistency in assessment across interview panels. The mean interpanel ICCs for the total score, the overall impressions, and the individual interviewer rank orders are presented in Table 2. Again, the more global scoring systems were more reliable. Using the Spearman-Brown prophecy formula, it was determined that two panels observing each candidate would be sufficient to obtain an interpanel reliability of 0.80 for the ranking of candidates. There were no significant differences in the reliability of selection processes between programs that utilized structured versus unstructured interviews.

#### *Consistency of candidate ratings between programs*

To calculate the reliability of the selection process between orthopedic programs, pairwise program-by-program correlations were calculated for total scores, overall scores, individual raters' final ranks, and the consensus final rank orders (see Table 2). The low correlations seen (ranging from 0.14 to 0.18) indicate that there is considerable variation between programs in the generation of rank orders for candidates.

#### *Changes in interviewer opinion over the process*

Finally, for each program, a correlation was calculated between the mean rank order for candidates across interviewers within a program and the consensus final rank orders of candidates generated by that program. The mean of these correlations across all programs was 0.88 (SD 0.10), suggesting that the discussion that occurred between selection committee members at the end of the interview process did not substantially alter the final candidate rank orders generated by each independent interviewer assessment.

#### **Comments**

Tarico and colleagues [4,5] developed the critical incident technique of resident selection in order to standardize the interview process. This study illustrated the use of the critical incident technique to define the characteristics and critical behavioral descriptors of candidates that are essential for successful resident performance in orthopaedic sur-

gery. Using this information a rating form was developed by which assessments could be made of the resident selection processes across 13 medical programs in Canada. This allowed for a standardized method of statistically analyzing and comparing each program in this study.

Several interesting results arose from the use of this opportunity. First, selection committee members across a wide variety of programs and geographic locations consistently indicated similar important qualities and behavioral descriptors used for the selection of candidates to orthopedic residency programs. Thus, it was relatively easy to create a 9-item form that satisfied members from every program involved in the study. The internal consistency of the 9-item scale (a mean of 0.71) and the consistently moderate correlations between each item and the final ranking generated by a given faculty member (means of 0.34 to 0.50) suggest that items were used as a coherent scale by interviewers and captured well their overall opinion of the candidates.

Second, there was quite reasonable agreement both in the scores generated by different interviewers participating in the same interview, and in the scores generated by different panels of interviewers participating in different interviews for the same program. The estimate that only two interviewers are needed per panel and two interview panels are needed per program in order to obtain acceptable reliability of scores speaks well of an interview process that includes the use of our rating form. Further, the scores generated using the form correlated well with the final rankings generated by a given program after all candidates had been interviewed. The scale, therefore, appears to have quite reasonable reliability as an evaluation of candidate performance, and impressive validity as a measure of the program's opinion of the candidates.

The results of this study indicate that, using the rating form generated in this study, it is not necessary for selection committees to use large numbers of interviewers in order to generate high reliabilities and consistent assessments of candidates. This information could be used by programs to alter interview practices aimed at a reducing the time and resources allocated to the selection of surgical residents each academic year, while maintaining an acceptable level of reliability within their selection process. It is felt by some that a large interview panel will potentially allow a greater ability to explore a candidate's personality, history, and

experience while minimizing the possibility of undue influence from one panel member's personal bias regarding an individual applicant [6]. Smaller interview panels, on the other hand, have been purported to allow for better interaction among panel members and to lend themselves to a more relaxed and informal atmosphere [7]. Either argument may be considered legitimate, and the particular approach taken is likely to depend on the particular program. The current results suggest, however, that these considerations should not necessarily be driven by the need for reliability. A small number of individuals per interview combined with a small number of interviews appears to have sufficient reliability to perform the task.

Similarly, it has been suggested that one potential problem with interviews is the difficulty of comparing candidates if the interview process changes with each candidate. Interviews can be either unstructured, semistructured, or structured [6]. A structured interview is one that utilizes standardized questions for all applicants, provides sample answers for the purpose of evaluation, and uses a panel for the interview. Increasing interview structure allows interviewers to standardize questions in order that all candidates receive questions of the same content and level of difficulty. It has long been reported that the reliability and validity of interviewing candidates improves as structure is added [8,9]. Similarly, a review of the literature indicates that there is a lack of reliability and validity with unstructured interviews [10–12]. In our study 33% of programs (4 of 12) used some form of interview structure, and 67% (8 of 12) used unstructured interviews. There was no significant difference in the reliability of selection processes between programs that utilized structured versus unstructured interviews. Again, based on our data it would appear that the decision by a program to use structured or unstructured interviews need not be driven by the need for reliability but rather is a matter of preference within the program.

However, it is worth noting that across programs, ratings were very unreliable. That is, there was essentially no relationship between the scores that candidates received at one program and the scores that they received at a second program. This may not be surprising with respect to resident ranks, as many programs may have different sets of priorities for the selection of candidates. It was more surprising, however, to find that the scores on the scale themselves were very different across programs. One possible reason for this variability across programs is that, while most programs agree in the abstract what the important dimensions of a good resident might be, each program's perception of how they value these dimensions or in fact what these

dimensions look like in practice may be quite different. This, however, is only speculation and requires further research to properly understand what this variability across programs really means.

The accurate selection of residents is becoming increasingly important due to heightening economic and societal forces. Committee members often dedicate a great deal of time and effort to resident selection. The ability to distinguish excellent from poor candidates by undergoing a systematic process of resident selection in a consistent manner is important for residency program selection committees. A greater understanding of the factors involved in the selection of residents may help programs develop consistency from year to year in making decisions based on the information available at the time of selection. Certainly the technique used in this study affords programs the opportunity to critically evaluate the process of resident selection at an institution or series of institutions, review the areas of desired emphasis during resident selection, and to evaluate whether their desired focuses and objectives are being met. This technique is not only useful for the purposes of orthopedic residency selection, but also offers a manner of evaluating selection procedures in other surgical and medical disciplines.

## References

- [1] Provan J, Cuttress L. Preferences of program directors for evaluation of candidates for postgraduate training. *Can Med Assoc J* 1995; 153(7):919–23.
- [2] Wagoner NE, Suriano JR, Stoner JA. Factors used by program directors to select residents. *J Med Educ* 1986;61:10–21.
- [3] Clarke J, Wigton R. Development of an objective rating system for residency applications. *Surgery* 1984;96:302–7.
- [4] Tarico V, Smith W, Altmaier E, Franken E. Critical incident interviewing in evaluation of resident performance. *Radiology* 1984;152: 327–9.
- [5] Tarico V, Smith W, Altmaier E, Franken E, Berbaum K. Development and validation of an accomplishment interview for radiology residents. *J Med Educ* 1986;61:845–7.
- [6] Edwards J, Johnson E, Molitor J. The interview in the admission process. *Acad Med* 1990;65:167–77.
- [7] Vickers M, Reeve P. Selecting specialists for training: 2. *Br J Hosp Med* 1993;50(11):663–9.
- [8] Wagner R. The employment interview: a critical summary. *Pers Psychol* 1949;2:17–46.
- [9] Wright OR. Summary of research on the selection interview since 1964. *Pers Psychol* 1969;22:391–413.
- [10] Grant DL. Issues in personnel selection. *Profess Psychol* 1983;11: 103–7.
- [11] Hakel M. Interviews. *Annu Rev Psychol* 1986;37:351–80.
- [12] Tenopyr ML, Oeltjen PD. Selection and classification results. *Annu Rev Psychol* 1982;33:67–97.