

## Scoring of Orthopaedic Residency Applicants: Is a Scoring System Reliable?

*Douglas R. Dirschl, MD*

The purpose of the current study was to examine the interobserver reliability of a scoring system designed to objectify the screening of orthopaedic residency applications. Forty residency applications were selected randomly from those received in 1998 and were scored independently by six observers. The scoring system used included objective and subjective elements. Weights were assigned to individual data elements based on the results of a previous investigation. Interobserver reliability was calculated using an intraclass correlation coefficient. The overall reliability of the scoring system revealed an intraclass correlation coefficient of 0.80. The intraclass correlation coefficients for individual elements in the scoring system ranged from 0.28 to 0.98. The intraclass correlation coefficient was high for elements that were numeric, but was low for elements that were more subjective. Observers thought the scoring form was easy to complete but found some of the subjective data elements difficult to score. No benchmarks exist by which to define an acceptable intraclass correlation coefficient of a scoring model for resident applicants. Even with the use of careful definitions, raters had poor reliability in scoring elements such as letters of recommendation and personal statements. The results of the current study indicate that one score of objective data elements is

adequate to screen a residency application for these elements. Programs screening applicants based on subjective elements, however, should be aware that great interobserver variability exists in the interpretation of those elements.

Selection of residents is one of the most difficult tasks done by medical school faculty, and is a task for which few faculty are well prepared. Each residency training program must have a method for determining which applicants it most desires to recruit. Not unexpectedly, these methods vary widely among programs, with some programs even using computer software to screen applications for numerous variables.<sup>2</sup> Although a structured approach to screening and interviewing resident applicants has been advocated by some,<sup>2,3</sup> screening applicants' files continues to be done in most programs on a relatively subjective basis by numerous members of an admissions committee.

Each training program should consider assessing its own resident selection methods to determine how well the methods identify candidates who will meet or exceed the expectations of the training program. The ideal methodology would be reliable, reproducible, and strongly predictive of a positive outcome (acceptable or better performance in the training program). The current study investigated a portion of the process of evaluating residency applications by determining the interobserver re-

---

From the University of Oregon Health Sciences Center, Portland, OR.

Reprint requests to Douglas R. Dirschl, MD, University of Oregon Health Sciences Center, 3181 SW Sam Jackson Park Road, Portland, OR 97201.

liability of a specific scoring system designed to objectify the screening of residency applications in one orthopaedic training program.

**METHODS**

A scoring system for screening residency applications was created by the author, using data historically collected and thought to be important by the department of orthopaedics at the author's institution. The scoring system included objective elements and subjective elements (Table 1).

Objective elements were those for which one numerical calculation could be made by the observer reviewing the applications. These included the number of honors grades in the basic science and clinical years of medical school, whether the applicant had achieved membership in Alpha Omega Alpha, the percentile score on the United States Medical Licensing Exam-

ination, Part I, the number of research projects done while in medical school, the number of abstracts or manuscripts published, and whether the applicant's medical school ranked in the top 10 nationally.<sup>1</sup>

Subjective elements were those that involved much more interpretation by the reviewer. These included determination of the numbers of activities volunteered by the applicants that involved the use of gross motor and fine motor skills, the number of leadership or volunteer activities, an evaluation of three letters of recommendation, and evaluation of a personal statement.

Weights were assigned to the individual data elements in the scoring system by the author based on the results of a previous investigation into the correlation between information in the residency application and subsequent performance as a resident.<sup>3</sup> Scores were normalized to a 100-point scale.

Forty orthopaedic residency applications were selected at random from those received by the de-

**TABLE 1. Scoring System for Screening Resident Applications**

	<b>Points</b>	
<b>Academics (15 points)</b>		
Number of honors medical school Years 1 and 2 (Anatomy, Physiology, Microbiology, Pathology, Biochemistry, Pharmacology)	(0=0-2; 1=3-4; 2=5-6)	___
Number of honors medical school Years 3 and 4 (Surgery, Medicine, Obstetrics and Gynecology, Pediatrics, Psychiatry)	(0=0; 1=1-2; 2=3; 4=4; 5=5)	___
United States Medical Licensing Examination Part I score (percentile)	0= < 60%; 1=60%-80%; 2= > 80%)	___
Alpha Omega Alpha	(0=no; 1=yes)	___
Research projects	(0=0-2; 1=3-5; 2= >5)	___
Abstracts and publications (accepted or published)	(0=none; 1=1-2; 2=3 or more)	___
Medical school reputation	(0=good; 1=top 10)	___
<b>Skills (3 points)</b>		
Gross motor (athletics, carpentry)	(0=0-2; 1= > 2)	___
Fine motor (music, woodworking)	(0=0-2; 1= > 2)	___
Leadership and volunteer activities	(0=0-3; 1=4 or more)	___
<b>Subjective (7 points)</b>		
Letters of recommendation (3)	(0=average; 1= > above average; 2=superior)	
	Letter 1	___
	Letter 2	___
	Letter 3	___
Personal statement	(0=average; 1=above average)	___
<b>Subtotal score (of 25 points)</b>		
<b>Bonus points (4 points)</b>		
Intangibles and accomplishments (rocket scientist, elite athlete, externship performance)	(0=average; 1=above average; 2=excellent; 3=outstanding; 4=stupendous)	___
<b>Total score (add bonus to subtotal score)</b>		
<b>Final score (total score x4)</b>		

partment of orthopaedics at the author's institution through the Electronic Residency Application Service for the 1999 National Residency Matching Program. Of the 40 applications selected, eight (20%) were from women. Because the Electronic Residency Application Service does not provide photographs or information about ethnicity at the time of screening of applications, this information was not available to observers who were scoring the applications. The 40 applications were reviewed and scored independently by six observers using the scoring system. Observers included five orthopaedic faculty members of the resident selection committee (two assistant professors, two associate professors, and one professor) and the orthopaedic residency coordinator (a nonphysician). Application of the scoring system to the 40 applications was for research only; the scores resulting from this study were not shared with the observers or the remainder of the resident selection committee. The resident selection committee was unaware of which applicants had been part of the study, and the data generated in the study were not used for decision-making in the 1999 National Residency Matching Program.

Interobserver reliability was calculated for each data element and for the overall score using an intraclass correlation coefficient. The intraclass correlation coefficient, which reflects the degree of correspondence and agreement among raters, is considered excellent when above 0.90, good when above 0.75, and poor when below 0.75.<sup>4</sup> It has been suggested that an intraclass correlation coefficient greater than 0.90 indicates a clinically useful test.<sup>4</sup>

## RESULTS

The overall interobserver reliability for the scoring system revealed an intraclass correlation coefficient of 0.80. The intraclass correlation coefficients for the various elements making up the scoring system ranged from 0.28 to 0.98 (Table 2). In general, the intraclass correlation coefficient was high for elements that were numeric and fairly objective (Alpha Omega Alpha, 0.98; number of publications, 0.89; number of honors grades in Years 1 and 2, 0.84), whereas it was low for data elements that were much more subjective (number of activities involving fine motor skills, 0.28; quality of personal statement, 0.30; quality of letters of recommendation, 0.50). Only election to Alpha Omega Alpha (0.98) and United States Medical Licensing Examination, Part I score (0.95) reached excellent reliability, whereas number of publications (0.89), number of honors grades in the basic science (0.84) and clinical years (0.83), and overall score (0.80) revealed good reliability. All of the remaining elements had poor interobserver reliability. There was no significant difference in the overall scores of the female applicants and the male applicants ( $p > 0.5$ ).

Observers generally thought the scoring form was easy to complete, although three observers commented that some of the data ele-

**TABLE 2. Intraclass Correlation Coefficients**

<b>FINAL SCORE</b>	0.80
<b>Objective Data Elements</b>	
Number of honors in medical school Years 1 and 2	0.82
Number of honors in medical school Years 3 and 4	0.84
Score on United States Medical Licensing Examination Part I	0.95
Elected to Alpha Omega Alpha	0.98
Number of research projects	0.72
Number of publications	0.89
<b>Subjective Data Elements</b>	
Medical school reputation	0.62
Number of activities using gross motor skills	0.30
Number of activities using fine motor skills	0.28
Number of leadership and volunteer activities	0.54
Quality of letters of recommendation	0.50
Quality of personal statement	0.38
Intangibles	0.52

ments scored were unnecessary (number of research projects, gross motor skills) and that some of the more subjective data elements were difficult to score. There was no significant difference in intraclass correlation coefficient based on academic rank ( $p > 0.5$ ); in addition, the scores of the residency coordinator did not differ significantly from those of the attending physicians ( $p > 0.5$ ).

## DISCUSSION

Screening of residency applications can be a daunting process; it is not unusual for an orthopaedic training program to receive 600 applications in 1 year. Some methodology must be used to cull these applications down to approximately 60 applicants who can be invited to interview with the training program. Methods and data used to screen the applications vary widely among programs. In general, all programs seem to value hard data, but family medicine programs seem to find the dean's letter and personal statement more valuable than do surgical training programs.<sup>5</sup> Some have contended that, as training programs become more competitive (have large numbers of applicants for each available residency), they rely more heavily on academic credentials as a means of screening applicants and less heavily on items such as a dean's letter or personal statement.<sup>6</sup> One anesthesiology training program has incorporated computer software in its screening methods for residency applicants.<sup>2</sup>

Although the overall intraclass correlation coefficient for the scoring model used in the current study is good, it does not approach the level necessary for a reliable clinical test. Because this study is the first of its type, no benchmarks exist by which to determine the acceptability of a scoring model for residency applicants based on the intraclass correlation coefficient. The appropriate value of the intraclass correlation coefficient to indicate a reliable scoring model for residency applicants is not known. Reliability in scoring the objective elements in the model generally was excellent, whereas reliability was universally poor in

scoring the subjective elements. Even with the use of careful definitions and categoric criteria for assessment, the raters had generally poor interobserver reliability in the scoring of data elements such as letters of recommendation, personal statement, and number of activities involving the use of fine or gross motor skills.

The finding in this study that objective, numerical data can be assessed reliably and that subjective data cannot be assessed reliably is no surprise; nor is it surprising that combining objective and subjective data into one model results in intermediate overall reliability. Overall interobserver reliability in this study could have been modified if more or less weight were given to objective and subjective elements. The weights in this study were selected based on the results of a previous study, but it is the task of each residency program to determine the mix of objective and subjective data it wishes to use in screening applicants to its training program. This study reports only the interobserver reliability of one screening process for applicants and does not attempt to pass judgment as to whether objective or subjective data are a better predictor of performance during residency. One interesting finding in this study was that a trained clerical person could screen applicants as effectively and reliably as the faculty.

The results of the current study indicate that one scoring of objective data elements, even by a nonphysician, is adequate to screen a residency application for these elements. If, however, a program desires to screen based on subjective elements, it should be aware that great interobserver variability exists in the interpretation of those elements. If residency training programs think that subjective data elements are important to the process of residency selection, very specific and strict identification and definition of each specific data element used will be necessary for such elements to be used reliably in the residency selection process. Perhaps this sort of specific and strict identification and definition of data elements important to training programs is a logical next step for programs wishing to improve the reliability of their

resident selection process. If a program can determine what data elements it values most highly in selecting residents, perhaps additional study can reveal a more reliable method for evaluating those specific subjective data elements. Continued assessment of the effectiveness of the residency selection process may be beneficial to all training programs.

### References

1. Anonymous: 1998 Annual Guide: Best Graduate Schools. US News and World Report, March 1998.
2. Baker JD, Wallace CT, Cooke JE, Alpert CC, Acklerly JA: Selection of anesthesiology residents. *South Med J* 80:1031-1035, 1987.
3. Dirschl DR, Dahners LE, Adams GL, Crouch JH, Wilson FC: Correlating selection criteria with subsequent performance as residents. *Clin Orthop* 399:265-271, 2002.
4. Portney LG, Watkins MP: *Foundation of Clinical Research: Application to Practice*. East Norwalk, CT, Appleton and Lange 1993.
5. Taylor CA, Weinstein L, Mayhew HE: The process of resident selection: A view from the residency director's desk. *Obstet Gynecol* 85:299-303, 1995.
6. Wagoner NE, Suriano JR, Stoner JA: Factors used by program directors to select residents. *J Med Educ* 61:10-21, 1986.