

Resident Selection and Predictors of Performance

Can We Be Evidence Based?

Douglas R. Dirschl, MD; Edmund R. Champion, MD; and Karen Gilliam

Selection of orthopaedic residents can be a difficult process; we have endeavored to make it more objective by developing a scoring methodology for screening applications. The purpose of this investigation is to determine if an academic score, using objective elements only, will discriminate among applicants and will correlate with outcomes. Applications to our orthopaedic residency program for 2004 and 2005 were assigned an academic score as a screening tool in the residency selection process. Data was analyzed for the entire group both by gender and whether the applicant had interviewed for the program. Additionally, the applications of program graduates over the past 5 years were retrospectively assigned academic scores, which were compared with outcomes of the training program. Academic scores for applicants formed a generally normal distribution, and residents training in the program generally had higher scores. The distribution of scores for female applicants was similar to male applicants; however, a greater percentage of female applicants interviewed for the program. Scores on the OITE and ABOS examinations tended to parallel academic scores, but faculty ratings of performance in the program showed no difference between those with high and low academic scores. Calculating academic scores makes the application screening process more objective but does not appear to correlate with outcomes of the training program.

Selection among and ranking of resident candidates is a task all orthopaedic training programs must accomplish and for which almost no orthopaedic academicians have had formal training. Methods for accomplishing this vary,

From the Department of Orthopaedics, University of North Carolina School of Medicine, Chapel Hill, NC.

Each author certifies that he or she has no commercial associations (eg, consultancies, stock ownership, equity interest, patent/licensing arrangements, etc) that might pose a conflict of interest in connection with the submitted article.

Each author certifies that his or her institution has waived approval for the human protocol for this investigation and that all investigations were conducted in conformity with ethical principles of research.

Correspondence to: Douglas R. Dirschl, MD, UNC Department of Orthopaedics, 3147 Bioinformatics, CB#7055, Chapel Hill, NC 27599-7055. Phone: 919-966-9072; Fax: 919-966-6730; E-mail: dirschld@med.unc.edu. DOI: 10.1097/01.blo.0000224036.46721.d6

with some programs using a scoring system, some a subjective analysis and committee discussion, and others a single chairman's or program director's assessment in making their decisions. Whatever the method, it is clear excellent people are matched into orthopaedic residency each year. But it is probably also clear, in this subjective process, excellent individuals are missed. At our institution, we have endeavored to make the selection method as evidenced-based and reliable as possible.

Our philosophy is to screen applications, but to rank people. Applications are screened based on data and objective information in the Electronic Residency Application Service (ERAS) application. We have endeavored over the years to develop a means by which we can assure our assessment of the data and information is as objective and reproducible as possible. We also believe, however, when it comes time to develop our rank list of residency applicants, we rank people according to our assessment of their personal qualities based on all the information we have at hand. We do not believe we should rank candidates by their credentials.

We have undertaken a series of investigations to attempt to address these issues. We hoped to better understand our own resident selection process from an evidence-based perspective. In the first study, a retrospective review, we determined which criteria in the residency application had the highest correlation with the subsequent performance of orthopaedic residents.³ We hoped this information would lead to the development of a scoring system that would objectify the screening of residency applications.

Summary of Previous Work and Development of "Academic Score"

The application files of 58 residents, who had completed our program over a 15 year period, were included. Application data collected included all the information traditionally collected in the residency application process at our institution. Measures of outcome included performance on

standardized examinations such as the Orthopaedic In-training Examination (OITE) and the American Board of Orthopaedic Surgery Part I examination (ABOS-1), as well as faculty performed ratings of resident performance on a 5-point Likert scale.³ The results of this study indicated faculty ratings had remarkable interobserver reliability. Faculty ratings correlated with the number of honors grades achieved in the third year clerkships during medical school ($p < 0.05$), but not with OITE and ABOS-1 scores. Letters of recommendation did not correlate with any of the outcome measures. It was concluded academic performance in core clinical clerkships in the third year of medical school seemed to be the best predictor of overall performance in the training program. It also appeared scores on standardized examinations had little correlation with predictor variables or faculty rating.³

Because the philosophy of ranking people rather than credentials could potentially rely heavily on personal recommendations, it was surprising to learn letters of recommendation did not correlate with outcomes. We postulated variability in interpretation of letters of recommendation by various individuals might explain this poor correlation, and undertook a second study to test the interobserver reliability in evaluating letters of recommendation.² One hundred seventy-four letters of recommendation were extracted from residency applications, were blinded to eliminate the identity of the applicant, the identity of the author, and the identity of the institution from which the letter came, and were rated by faculty observers.² The data suggested extremely low interobserver reliability in rating letters. It was concluded programs should be aware of the variability that exists in interpreting letters of recommendation, and this variability may be a serious limitation in the use of letters of recommendation as a good screening tool for residency applications.²

Further thought as to why correlations were poor between many of the predictor variables and the outcome variables led us to consider the concept of range restriction. Range restriction is a well-known statistical concept that occurs when an observer's view of the population variable measured is skewed because the examined sample represents only a small portion of the normal distribution of the entire population. When, for example, one tail of a normal distribution is sampled, statistical evidence would indicate good reliability and good regression models cannot be obtained. We believed this concept to apply to our situation, because the applications included in our studies were those of residents who had been admitted to our training program and represented largely the upper portion of a normal distribution of the hundreds of applicants we had for the program each year.

It followed from this realization that an appropriately constructed scoring system for residency applicants might

perform with better reliability if it was applied to all applicants to an orthopaedic training program, rather than just to the individuals who matched and trained in the program. With this in mind, we undertook another study to determine the interobserver reliability of a specific scoring system designed to objectify the screening of residency applications.¹ The system included data elements weighted according to the results of our previous investigation (Table 1). The objective elements (elements 1–7 in Table 1) were later defined as the “academic score” (Table 2).

The scoring system was tested by selecting 40 applications from those received at our institution.¹ The selected applications were reviewed and scored by six observers, five of whom were orthopaedic faculty and one of whom was a nonphysician residency coordinator. The results (Table 1) indicated the intraclass correlation coefficients (ICC) for the components of the academic score were generally quite high. An ICC of greater than 0.8 was considered extremely good. Results for the subjective elements indicated the intraclass correlation coefficient for these elements was very poor (range, 0.2–0.36),⁴ indicating raters did not agree in their assessment of these data elements. There were excellent correlations between the scores of faculty and those of the nonphysician on the academic score elements. We concluded a single scoring of objective data elements (the academic score), even by a nonphysician, would be adequate to reliably screen a residency application. We also concluded great variability exists in the interpretation and scoring up subjective elements, and these should probably not be used in screening applications.¹

TABLE 1. Intraclass Correlation Coefficient of Scoring Elements

Variable	ICC
Academic score elements	
Number of honors grades in medical school years 1 and 2	0.82
Number of honors grades in medical school years 3 and 4	0.84
Score on USMLE Part I examination	0.95
Elected to Alpha Omega Alpha	0.98
Number of research projects	0.72
Number of publications (abstract or manuscript)	0.89
Subjective data elements	
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

ICC = intraclass correlation coefficient; USMLE = United States Medical Licensing Examination

TABLE 2. Components of the Academic Score

Variable	Academic Score
Number of honors grades medical school years	0 = 0–2
1 and 2 (anatomy, physiology, microbiology, pathology, biochemistry, pharmacology)	1 = 3–4
Number of honors grades medical school years	2 = 5–6
3 and 4 (surgery, medicine, obstetrics and gynecology, pediatrics, psychiatry)	0 = 0
	1 = 1–2
	2 = 3
	4 = 4
	5 = 5
USMLE Part I Examination (percentile score)	0 = < 60th
	1 = 60th–80th
	2 = > 80th
Elected to Alpha Omega Alpha	0 = no
	1 = yes
Number of research projects	0 = 0–2
	1 = 3–5
	2 = > 5
Publications (abstracts and manuscripts)	0 = 0
	1 = 1–2
	2 = 3 or more
Medical school reputation (top 10 in US News and World Report)	0 = no
	1 = yes

USMLE = United States Medical Licensing Examination

Current Study on the Application of the Scoring System

Having shown a scoring system based on objective data elements (academic score) could perform with high interobserver reliability, we put the system into use for screening application in our orthopaedic residency program. Having used this system for screening applications in the 2004 and 2005 NRMP matches, we believed we had sufficient data to answer some questions about the performance of the scoring system. The purpose of the present investigation is to determine if, using objective data elements only, the academic score will discriminate among applicants and whether it will correlate with outcomes.

MATERIAL AND METHODS

We included 1006 completed applications to our orthopaedic residency program for the NRMP in 2004 and 2005. Each application had an academic score assigned by the residency coordinator and verified by the department chair. Academic scores could range from 0 to 15 (Table 2). Data were segmented by gender and by whether the applicant had interviewed for the program. Additionally, the applications of our residency program graduates over the past 5 years were retrospectively reviewed ($n = 20$). For each of these graduates, academic scores were calculated based on information from their residency application. Outcome data was also collected for this group, and included: (1) mean percentile score on the OITE over the 4 years of orthopaedic training; (2) percentile score on the ABOS-1 ex-

amination; and (3) rating of overall performance by the chair and program director on a 5-point Likert scale. Where appropriate, statistical analysis was conducted by segmenting the data into categories and applying a Chi Squared analysis.

RESULTS

The academic scores formed a generally normal distribution (Fig 1). We cannot explain the large number of applicants with an academic score of 1; review of the data, however, indicates this low score was not given erroneously. The vertical line in the Figure 1 indicates the threshold above which applicants were generally invited to interview with our program. This threshold was not rigidly adhered to, however; there were, in individual cases, considerations other than academic score used in choosing whether to interview an applicant. The normal distribution of scores indicates the scoring system discriminates reasonably well between applicants, although it does not provide any information about the relative importance of the items being screened for.

While the program generally tended to interview candidates with higher scores, it did not adhere rigidly to the academic score as the only means of choosing applicants to interview (Fig 2). Applicants with academic scores lower than the threshold may receive an interview when an individual known to and trusted by the program makes a strong personal plea for careful consideration of an applicant who would otherwise be rejected. However, even among applicants with the highest academic scores, not all were interviewed (Fig 2). Some of these applicants were invited to interview and declined the invitation, but some were not invited to interview. This is another indication the program considered factors other than academic scores, even in the applicants with the strongest academic performance.

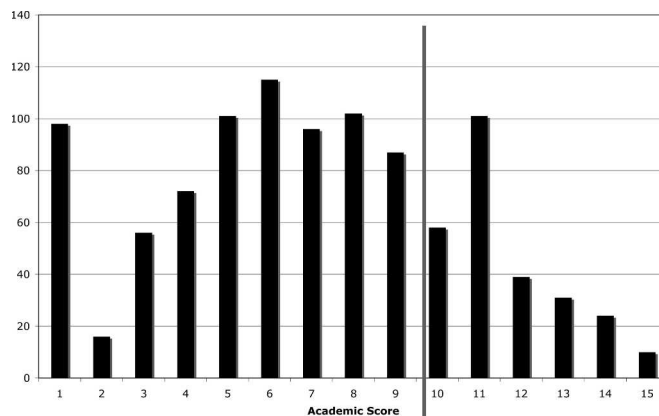


Fig 1. The academic scores for all residency applicants in 2004 and 2005 form a generally normal distribution.

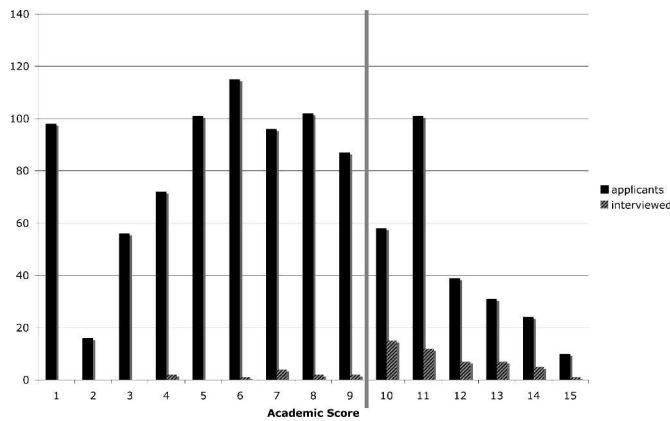


Fig 2. Applicants who are interviewed had generally higher academic scores than did the overall applicant pool.

Although female applicants represented only 12.5% (n = 126) of the total applicants, the distribution of the academic scores of female applicants was similar to that for male applicants (Fig 3). From among the entire pool we interviewed a higher percentage (p = 0.03) of female applicants (11%, n = 14) than male applicants (5%, n = 44). A greater (p = 0.07) proportion of interviewed female applicants had an academic score below the general threshold of 10 (4 of 14, or 28%) than was the case for interviewed male applicants (7 of 44, or 16%).

Residents completing the UNC orthopaedic training program in the last five years represented the upper portion of the distribution of applicant scores (Fig 4). However, our program is not solely populated by those applicants with the highest academic scores; there are a number of residents with scores below the threshold value, which can be explained by the two ways we rank applicants. An

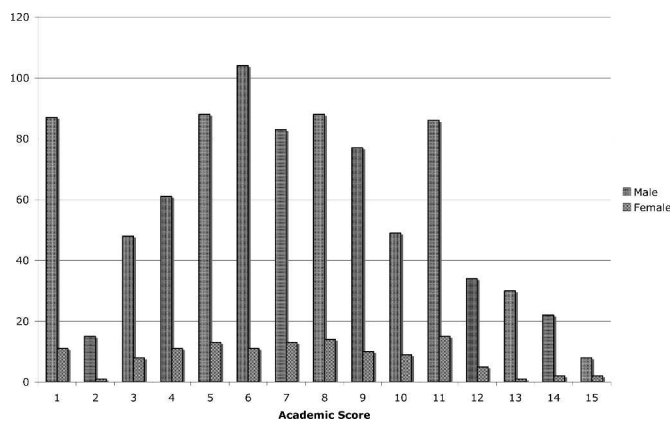


Fig 3. The distribution of academic scores for female applicants was not different than that of male applicants.

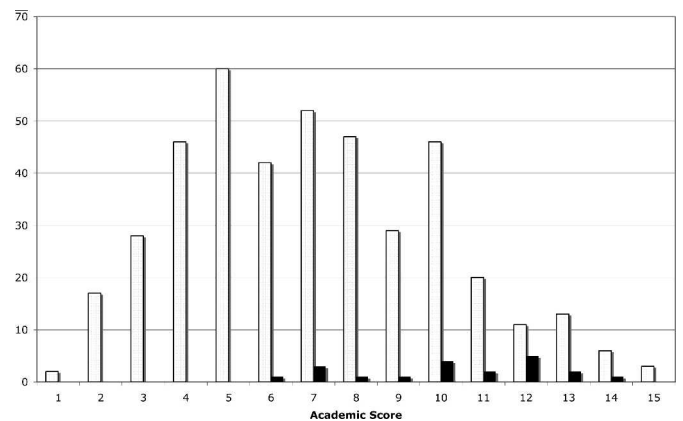


Fig 4. Residents training in the program represented generally the upper portion of the distribution of academic scores for all residency applicants.

applicant is offered an interview after an evaluation of his/her application materials, largely based on high academic score, or the program has a deep personal knowledge of the applicant. Our medical students and students from other schools who do rotations at our institution are better known by our program faculty because they have worked closely over a period of time. These applicants, in our program's selection process, automatically advance to the pool of applicants who are ranked, regardless of the academic scores on their applications.

Residents who match in the program by virtue of their application had higher academic scores than those that match by virtue of their rotation (p = 0.02) (Fig 5). This follows from the program's philosophy of screening applications, but ranking people. This latter group has aca-

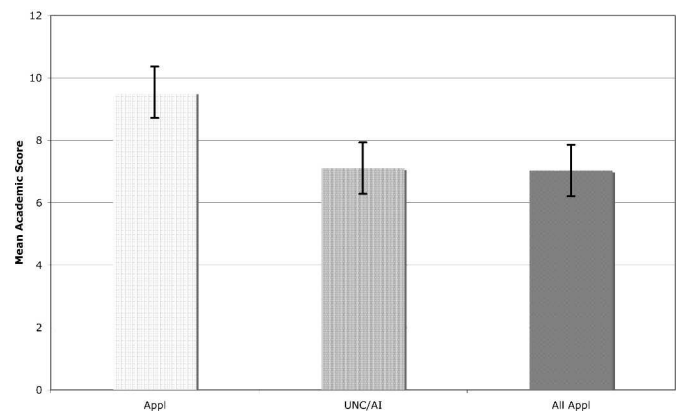


Fig 5. Mean academic scores for residents who matched into the training program by virtue of their academic score were higher than for those who matched by virtue of a rotation at our institution or for the entire applicant pool.

ademic scores essentially equal to those in the applicant pool at large. The academic scores of the two groups of residents had similar mean percentile scores on the OITE. We observed nearly identical result when comparing ABOS-1 scores in these two groups of residents. However, the faculty rating of overall performance in the residency program (1–5 on a Likert scale), was similar between these two groups of residents who matched by somewhat different paths (Fig 6). It remains unclear whether the faculty rating or the OITE or ABOS-1 score is a better measure of outcome, but only performance on the examinations correlates with the academic score.

DISCUSSION

Some might argue this line of investigation is nothing more than an academic exercise that has no practical value in the selection of residents. We believe there is useful information to be gained by an objective approach to the issue of resident selection and screening. We think it is important to better understand what our program values in its residents and how well the program achieves the outcomes it desires. It is important to continually assess the process of resident selection and to learn if we can be more effective at this difficult and important task.

Our scoring system for academic data appears to distinguish between and discriminate among applicants fairly reliably. It does not have a bias for or against female applicants. The data also confirm the fast track methods used for our students and students doing rotations in the department seem to give those students an advantage in the residency selection process. Students being interviewed and ranked through the screening pathway have higher

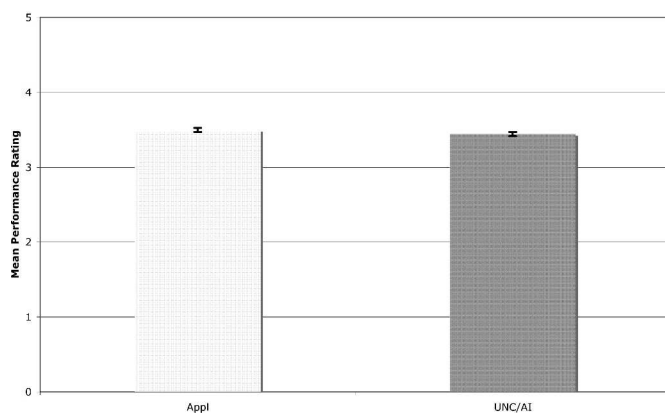


Fig 6. Mean faculty performance rating scores for residents who matched into the training program by virtue of their academic score were no different than for those who matched by virtue of a rotation at our institution.

academic scores than those included on the program's rank list through the other pathway. We also see there seems to be no difference in faculty performance ratings between residents regardless of their academic scores or whether they matched our program application or personal knowledge. The OITE scores and ABOS scores of residents, however, remain fairly consistent with their application academic score. Applying the scoring system enables us to see, despite a lower academic score threshold for acceptance into the program, that the residents who rotated in the department match their peers (who were admitted as a result of the screening process) in faculty ratings of performance.

Is this system objective? The data suggest the screening system for applications is objective, but the selection system for ranking applicants certainly is not. At this time, we are unable to determine whether our system selects for the outcome measure we desire because we must more clearly define and state exactly what those outcomes are. The current system retrospectively assesses resident performance and tries to match pre-existent characteristics with what it is thought will be their eventual abilities and professional qualities. Ideally, there would be an array of desired final targeted outcome measures and a similar group of selection criteria that if present would be predictive of the resident being very likely to achieve those specific outcomes. Currently, our selection methodologies are far from sociometric specificity, but these investigations are a step towards more clearly defining what outcomes we truly value and assist us in determining what pathways are most likely to lead us to achieving these outcomes.

There are a few issues related to our use of testing data in this study that may impact on the objectivity of our system. The USMLE Part I examination is a standardized national examination that returns for each examinee a three-digit score. Three-digit scores are equated across time and exam form, such that identical three-digit scores—regardless of the year in which the examination was taken—imply equivalent levels of performance. That said, it is possible (though not proven) USMLE test results can be biased by medical school characteristics and curricula designed to teach specifically to the format and content of the USMLE examination. Finally, the USMLE examinations are not designed to predict future performance of physicians.

The authors understand the OITE and ABOS examinations are not intended, nor have they been validated, for the use we have put them to in this study. The OITE does provide a percentile score for each resident compared to all residents in the same PGY year taking the examination. This percentile score, however, is not standardized nor statistically validated for comparisons to be made between

residents in different PGY years or over numerous years for the same resident. The ABOS examination is validated only at the level of “pass/fail”; the comparison of percentile scores on the ABOS examination has not been validated. In the case of both of these examinations, however, we believe the use of percentile scores—while not statistically validated—can provide valuable information to our program on the performance of our scoring methodology for screening resident applicants.

Finally, if we were to ask if our system misses individuals who would be very good orthopaedic residents, we must acknowledge any system for screening over 500 applications with the ultimate goal of matching four residents will most certainly miss some very good people. Much

broader evaluation of the entire applicant pool, however, is necessary to determine if any selection criteria can be highly successful in decreasing the likelihood a program will accept a truly poor applicant or the likelihood a truly excellent applicant is denied acceptance.

References

1. Dirschl DR. Scoring of orthopaedic residency applicants: is a scoring system reliable? *Clin Orthop Relat Res.* 2002;399:260–264.
2. Dirschl DR, Adams GL. Reliability in evaluating letters of recommendation. *Acad Med.* 2000;75:1029.
3. Dirschl DR, Dahners LE, Adams GL, Crouch JH, Wilson FC. Correlating selection criteria with subsequent performance as residents. *Clin Orthop Relat Res.* 2002;399:265–271.
4. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33:159–174.