PAPERS FROM THE APDS SPRING MEETING

To See Ourselves as Others See Us: Self-assessment in Surgical Residency

EDWARD M. KWASNIK, MD, AND JEANNE CARTER

Purpose: Although self-assessment is a widely used educational technique, the value of self-evaluation in surgical residency has not been clearly defined. This study was undertaken to assess the ability of residents to evaluate themselves using the same standards as the surgical faculty and to determine how this information may be used in a surgical training program.

Methods: Categorical surgical residents were asked to grade themselves (scale, 1–10) in 12 performance characteristics by completing the same forms used by the faculty for 3 quarterly evaluation periods. Mean faculty grades ± standard deviation were computed for each resident and compared with resident self-evaluation grades for a global rating scale and for specific performance characteristics. Comparisons were made by Pearson correlation analysis, analysis of variance, and Bonferroni's multiple comparisons test with significance accepted at $p < 0.05$.

Results: A significant correlation ($r = 0.47$; $p < 0.0098$) was identified between faculty and resident in the global score and in the specific performance characteristics of knowledge ($r = 0.51$; $p = 0.0052$), clinical judgement ($r = 0.48$; $p = 0082$), and technical ability ($r = 0.52$; $p = 0.0036$). For residents who accurately estimated their scores, mean faculty score was 7.3 ± 0.64, which was significantly higher than the mean score in overestimators (6.4 ± 0.89; $p < 0.05$), which in turn was significantly lower than the mean score in underestimators (8.2 ± 0.58; $p < 0.001$). No significant correlation was obtained for ethical standards and interpersonal relationships or between PGY level and accuracy of self-evaluation.

Conclusions: Whereas resident self-evaluations correlated positively with faculty ratings both globally and in certain performance categories, the majority of residents over- or underestimated their abilities. Information from such self-evaluations may be useful in counseling residents as well as in monitoring and improving the evaluation process. (Curr Surg 1999; 56:145–148. © 1999 by the Association of Program Directors in Surgery.)

"O wad some Power the giftie gie us
To see oursels as others see us"

—Robert Burns

Self-evaluation is a common educational and motivational tool both in the corporate world as well as in most of our elementary and secondary school systems.[1] It is interesting, however, that very little information is available regarding the application of this methodology in graduate medical education.[2–4] Over the past 2 decades, there has been but a single study published in the surgical literature[5] that compared resident self-evaluations to those performed by faculty and peers.

Given the great interest in self-assessment that exists in the broader educational community, this study was undertaken to investigate the ability of surgical residents to rate themselves in comparison with faculty both on a global scale and in specific performance categories. Additional focus in data analysis was placed on those residents who significantly over- or underestimated their abilities in relation to the teaching staff, as well as on the influence of PGY level on accuracy of self-assessment. Our ultimate goal was to define potential roles for self-assessment in a surgical residency program.

METHODS

Categorical surgical residents were asked to grade themselves by completing the same forms used by the faculty at the end of 3 separate quarterly evaluation periods between June 1997 and September 1998. The evaluation sheet was formatted as a graphic rating form in which 12 performance characteristics (Table 1) were graded on a scale from 1 to 10. This scale was further subdivided into numerical ranges that were anchored to specific labels (unsatisfactory, 1–3; satisfactory, 4–7; superior, 8–10). These categories were also defined by specific behavioral examples for each characteristic.

A total of 29 of a possible 30 (97%) faculty-resident assessment pairs were obtained. Faculty grades were tabulated and mean values (± standard deviation [SD]) computed for each resident both as a global score comprised of all the

Table 1 Resident performance characteristics

Knowledge	Ethical standards
Clinical judgment	Responsibility
OR performance	Professionalism
OR preparation	Staff relationships
Clinical skills	Patient interactions
Organization	Conferences

components listed in Table 1, as well as for the individual characteristics of knowledge, clinical judgment, operating room performance, ethical standards, and interpersonal relationships with staff. These values were then compared with resident self-evaluation scores by Pearson correlation analysis. Further analyses were then performed to determine relationships between accuracy of self-evaluation and overall performance, and effect of PGY level on accuracy of self-assessment. Comparisons between means were analyzed by 1-way analysis of variance (ANOVA) with Bonferroni's multiple comparisons posttest as appropriate using a statistical software package (GraphPad version 2.01; GraphPad Software, Inc). Statistical significance was accepted at $p < 0.05$.

RESULTS

When faculty–resident assessment pairs for overall global scores were analyzed (Fig. 1), a significant Pearson correlation coefficient was obtained. Correlation between resident and mean faculty scores for specific performance characteristics are presented with their corresponding p values in Table 2. Whereas significant correlation coefficients between resident and faculty scores were identified in the categories of knowledge, clinical judgment, and technical ability, no statistically valid correlation was recognized for interpersonal relationships or for ethical standards.

In order to evaluate the relationship between accuracy of resident self-assessment and overall rating, 3 groups were formed based on degree of difference between faculty and resident evaluation scores. In group 1, (n = 12, 41%) this difference was < 0.5 units, for group 2 (n = 11, 38%) the resident scores were at least 0.5 units greater than mean faculty score, and in group 3, (n = 6, 21%) the resident scores were at least 0.5 units less than mean faculty score.

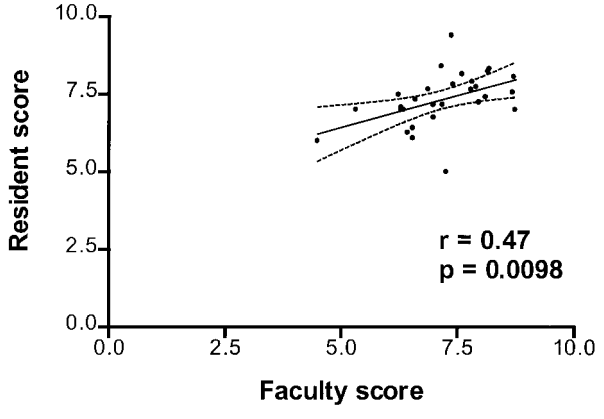

FIGURE 1. Correlation of resident and faculty global scores. Dotted lines = 95% confidence limits.

Table 2 Correlation data for specific performance characteristics

Characteristic	Correlation coefficient	p value
Knowledge	0.051	0.0052
Clinical judgment	0.48	0.0082
OR performance	0.52	0.0036
Ethical standards	0.31	NS
Interpersonal relationships with staff	0.21	NS

NS = not significant.

Mean faculty scores were determined for each group and are presented in Fig. 2. Initial analysis by 1-way ANOVA ($F = 12.2$; $p = 0.0002$) rejected the null hypothesis that group means were identical. Using Bonferroni's multiple comparisons posttest, significant differences were identified only between groups 1 and 2 ($p < 0.05$) and between groups 2 and 3 ($p < 0.001$). In this analysis, the mean score of residents in group 1 (7.3 ± 0.64) did not differ significantly from group 3 (8.2 ± 0.58), which was the only group to achieve a mean faculty grade commensurate with a superior. Whereas the other 2 groups fell into the satisfactory category, the only residents to be considered for or actually placed on probation were those who overestimated their abilities (group 2; mean score 6.4 ± 0.89).

In order to assess the influence of level of training and thus, indirectly, the effect of feedback from prior evaluations, resident faculty assessment pairs were assigned to 1 of the following classifications: junior (PGY 1 and 2), senior (PGY 3 and 4), and chief (PGY 5). Correlation coefficients and corresponding p values for these relationships are presented in Table 3. Whereas moderate positive correlation coefficients were obtained for the first 2 groups, a negative

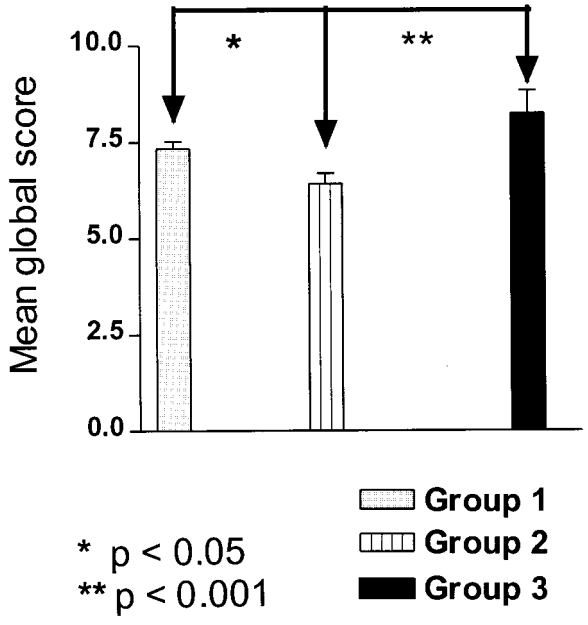

FIGURE 2. Mean global performance scores for resident groups based on degree of difference between resident and faculty scores. Group 1: difference < 0.5; group 2: resident scores at least 0.05 > faculty; group 3: resident scores at least 0.05 > faculty.

Table 3 Post level correlations

Postgraduate year	Correlation coefficient	p value
Chief (V)	−0.63	0.176 (NS)
Senior (III and IV)	0.42	0.194 (NS)
Junior (I and II)	0.57	0.0657 (NS)

NS = not significant.

coefficient was observed for the chief level. None of these results, however, reached statistical significance, possibly because of limitations of the statistical method in small samples.

DISCUSSION

The overall purpose of this study was to define better the role of self-assessment in a surgical residency program. Our results suggest that, despite certain limitations, self-evaluation may be applied with benefit to the development of residents both in terms of their global performance as well as in specific components of their surgical training. Furthermore, a review of data obtained from resident self-assessment may be useful for monitoring and improving the evaluation process itself.

In our study, a significant correlation was identified between faculty and resident ratings of global performance. The observed correlation coefficient of 0.47 was moderately strong, particularly when compared with a mean correlation coefficient of 0.35 between self and supervisor ratings, which was reported in a meta-analysis of 36 studies in the occupational psychology literature.[1] In addition, no significant correlation was found between self and teaching staff global scores in the only other such study[5] focused on surgical residents.

When the relationship between faculty and staff scores was analyzed in more detail, the ability of residents to evaluate their performance as compared with surgical staff varied considerably. The majority of our residents either over- or underestimated their capability as compared with the perceptions of the faculty. This may be in part because of differences in the way residents evaluate themselves or their peers as compared with the manner that most faculties approach the evaluation process. Specifically, several investigators[2–5] have found that factors that affected resident self-evaluation were complex, multidimensional, and included such traits as cognitive ability, interpersonal skills, and effort. Faculty, on the other hand, were influenced primarily by the interpersonal skills of the resident and secondarily by ability.

Residents in our program whose performance was thought to be marginal to the point of facing probation were less likely to identify their deficiencies. In this regard, our findings resemble those of Risucci et al,[5] who found highly significant negative correlation between the degree to which a resident self-rated too highly and ABSITE score, overall peer rating, and the overall supervisor rating. Such information is of obvious utility when counseling residents, because the first step in improving performance is the insight and realization on the part of the resident that certain areas of professional development are not proceeding at a satisfactory pace.

Of perhaps greater interest was the observation that our superior residents tended to grade themselves most critically. This phenomenon has been observed in a study of family practice residents,[2] in which it was thought that the perception of inadequacy led to harder work. Alternatively, in our study this cohort of residents may simply have benefited from halo effects based on isolated superior traits or personality characteristics.

When analyzing data regarding specific performance characteristics, significant correlation between resident and faculty scores was demonstrated in the areas of knowledge, clinical judgment, and technical ability. This may be attributed in part to the fact that these areas are most frequently assessed and feedback given on an almost daily basis during clinical interactions on the wards and in the operating room. In addition, such characteristics as knowledge may be linked to other more objective evaluation instruments, such as ABSITE scores and results of written quizzes.

In the category of ethical standards, no correlation was identified between resident and faculty evaluations. Judgments regarding character and personality are less objective and subject therefore to a higher degree of interrater variability, a factor that may have contributed to the low degree of correlation in these areas. Similar findings have been identified in studies comparing peer rating with those of faculty in 3 primary care specialties. Cognitive abilities were scored similarly by peers and supervisors, whereas faculty were more discriminating when rating interpersonal relationships.[6]

We have assumed, as do most programs, that faculty evaluations represent the gold standard for assessing resident performance.[7] A critical review of our data, however, demonstrated that the process is far from perfect. Several instances of rater errors, such as the halo effect and, more commonly, range restriction in both global and specific performance ratings, were observed. Identification of large discrepancies between faculty and resident assessment may call attention to situations in which significant process errors by the faculty may have impacted positively or negatively on the overall resident score. In these instances, the resident's self-assessment may indeed be more accurate and the proper response would be reeducation of the faculty regarding the pitfalls of the rating process.

A potential benefit of self-assessment is that by using the same global rating scale as faculty, residents will be made aware in detail of the criteria that are utilized in their summative evaluations. This process, as well as information from multiple feedback and counseling sessions, should lead to more accurate self-assessment during the later stages of training. Interestingly, our results demonstrated that when the entire group was subdivided based on years in the program, no significant correlation was identified at any level, possibly because of limitations of the statistical methods in small populations. Alternatively, more emphasis on discussing and resolving specific discrepancies between resident and faculty scores during formative feedback sessions may be needed to improve residents' self-assessment skills. In addition, evaluation criteria should be defined as objectively as possible to be most useful in this process.

The ultimate goal of self-assessment is to provide a basis for professional growth after the formal training period. Although objective measures of cognitive abilities (eg, SESAP)

are available and are very useful in this regard, to see ourselves globally as others see us, and to use that insight to direct personal and professional development, is a skill that should be fostered during residency training. In this regard, further investigation should be undertaken to relate self-assessment during residency with outcomes after training, such as board certification.

EDWARD M. KWASNIK, MD
JEANNE CARTER
Waterbury Hospital Health Center
Waterbury, Connecticut

REFERENCES

1. Harris MM, Schaubrock JA. Meta-analysis of self-supervisor, self peer and peer-supervisor ratings. *Personal Psychol* 1988;41:42–43.
2. Speechley M, Weston WW, Dickie GL, Orr V. Self-assessed competence: before and after residency. *Can Fam Physician* 1994;40: 459–464.
3. Kolm P, Verhulst SJ. Comparing self- and supervisor evaluations: a different view. *Eval Health Prof* 1987;10:80–89.
4. Wilson FC. Problems in the evaluation of surgical house officers by program directors. In: Lloyd JS, editor. *Residency Director's Role in Specialty Certification*. Chicago: American Board of Medical Specialties, 1985:35–41.
5. Risucci DA, Tortolani AJ, Ward RJ. Ratings of surgical residents by self, supervisor and peers. *Surg Gynecol Obstet* 1989;169:519–526.
6. Forsythe GB, McGaghie WC, Friedman CP. Construct validity of medical clinical competence measures: a multitrait-multimethod matrix study using confirmatory factor analysis. *Am Ed Res J* 1986; 23:315–336.
7. Gray JD. Global rating scales in residency education. *Acad Med* 1996;71(suppl):S55–S62.