Are the Communication and Professionalism Competencies the New Critical Values in a Resident’s Global Evaluation Process?

Mounir J. Haurani,* I. Rubinfeld,* S. Rao,* J. Beaubien,† J. L. Musial,* A. Parker,* C. Reickert,* A. Raafat,* and A. Shepard*

*Department of Surgery, Henry Ford Hospital, Detroit, Michigan; and †Human-Systems Integration Division, Aptima, Inc., Woburn, Massachusetts

BACKGROUND: The ACGME requires the assessment of resident competency in 6 domains. Global evaluations covering all 6 competencies are routinely used. Evaluators may be overly influenced by resident affability and availability, thereby resulting in a halo effect. We hypothesized that the Interpersonal Skills and Communications (ICS) and Professionalism (PR) competencies would unduly influence other competency scores.

METHODS: General surgery resident evaluations are performed by staff and peers on a rotational basis using competency-based questions. Each question is scored using a 5-point Likert scale. Mean individual composite scores for each competency were calculated and then correlated with other mean composite competency scores. Data from patient evaluations were similarly analyzed. A final correlation of competency scores to ABSITE scores, as an objective, standardized measure of a specific competency, Medical knowledge (MK) was also performed.

RESULTS: Results were available for 37 residents (PGY 1-5). There was a significant association between ICS scores and higher scores in MK (r = 0.52, p < 0.0004), PR (r = 0.826, p < 0.0001) and patient care (PC) (r = 0.619, p < 0.0001). No correlation, however, was found between patient evaluations of residents and their faculty/peer-based ICS scores. We found no association between ICS scores and improved patient evaluations. Lastly, we found no association between ICS or MK scores and ABSITE scores.

CONCLUSIONS: It was difficult to ascertain whether residents with better ICS scores had higher PR, PC, and MK scores because of the halo effect, improper completion of evaluations, or whether those residents were truly performing better clinically. External measures of resident performance did not correlate with faculty/peer evaluations of ICS and PR. Residency programs should consider adopting a more standardized way to objectively evaluate residents. (J Surg 64:351-356. © 2007 Association of Program Directors in Surgery. Published by Elsevier Inc. All rights reserved.)

KEY WORDS: resident training, evaluation, competencies, resident evaluation, communications competency, professionalism, halo effect, global evaluations

The Accreditation Council of Graduate Medical Education (ACGME) first created and then mandated use of the 6 competencies through its Outcomes Project. This initiative requires the assessment of resident competency in 6 domains with little guidance or recommendations for standardization of methods. Although the ACGME has clearly stated what the expectations are, it has not given surgical education programs more than suggestions for measuring progress and proficiency in these domains. Yet, the current expectation is that programs will not only measure resident performance in terms of these competencies, but also document steady progression and improvement.

Many strategies have been used to increase the reliability and utility of the resident evaluation process. A variety of methods have been described, including standardized examinations, oral examinations, objective standardized clinical examinations (OSCEs), and skills laboratories. The global evaluation has been a particularly popular assessment tool used nearly universally by a wide variety of residency programs. With this tool residents are reviewed at intervals (eg, monthly or by rotation) using a set of questions addressing specific competency-based behaviors and/or proficiencies. Responses are frequently scored on a 5-point or 10-point Likert scale to attempt standardization of results between subjects and evaluators. To accurately
evaluate a resident in these areas requires that reviewers be cognizant of the various subcomponents of each competency. This requirement can be met by fully educating evaluators about the competencies and/or by designing an assessment tool that encompasses all subcomponents of the competencies and clearly defines them. Faculty training is also necessary to ensure that responses are calibrated and reproducible between all faculty members; such training helps assure an acceptable inter-rater reliability necessary for evaluation validity.4,6 In reality, the logistics related to faculty training (ie, the educational process necessary to create and maintain such calibration) is a major barrier to its appropriate and widespread adoption.

Although valuable, faculty evaluations may not reveal a complete picture of resident performance. To provide more comprehensive assessment, many authorities have adapted a multi-source (360°) feedback evaluation, where all members of the health-care team evaluate resident performance. The concept of 360° evaluations was popularized in the business sector and has recently been accepted in medical education.7,9 Residents are evaluated by other residents at different levels of training, nurses, social workers, and even patients. The 360° evaluation process may be more consistent than other types of evaluations and is thought to provide better insight into the “softer” competencies such as professionalism (PR) and interpersonal and communication skills (ICS), which include interactions with peers, patients, and other members of the health-care team.10,11 Despite the added effort and paperwork involved in compiling 360° evaluations, they still may not provide additional information, especially if the different evaluator group scores are highly correlated.12 Ideally, these evaluations should also be completed by trained observers, but the logistics of training such a diverse group of evaluators is even more challenging than for faculty making this goal unrealistic.13

In an uncalibrated evaluation system, concern arises over the potential for assessment bias and likely sources of such bias. One of our particular concerns is that a halo effect may result from evaluators being unduly influenced by resident affability (ICS) and availability (PR). This concern is reflected in the well-accepted tradition of the “3 As of private practice”: affability, availability, and ability, which are the factors most important to a referring physician when picking a consultant in descending order of priority. Ample evidence exists that physicians lack insight into their biases toward other physicians. It has been shown, for example, that referral patterns are not based on physician skill but on relationships.14 A more recent study of breast cancer patients demonstrated that self-referred patients are more likely to go to high-volume, accredited centers.15 Despite such studies, physicians still report that consultant skill and competency are the most important factors when referring patients. The other factors that influence how patients are referred may be analogous to factors that could bias the resident evaluation process. To test the hypothesis that higher scores on measures of affability (ICS) and availability (PR) would result in higher scores on the other competencies, specifically ratings of medical

**METHODS**

The methods and research plan were reviewed by our institutional review board and found to be exempt. An internally developed rotational evaluation form based on the core competencies was used to evaluate residents. Each competency was divided into its essential components as defined by the ACGME, and questions were devised to assess residents’ performance in each component. The faculty evaluation of resident form uses a total of 22 questions, whereas the peer evaluation forms used only 9 or 11 questions (based on PGY-level—senior or junior resident). Behavioral anchors were created for all questions to assist in homogeneity and inter-rater reliability; these anchors are displayed with the question electronically while the evaluation is being completed. Responses are recorded on a 1-point to 5-point Likert scale. Individual scores marked N/A (not applicable) or missing were recoded as no data to avoid compromising mean calculations. Formal validity and inter-rater reliability for the questions have not been established. Ongoing, informal calibration has been pursued using routine communication between the program office and evaluating faculty as well as through monthly meetings of the departmental Education Committee. All evaluations were completed online using a departmental website evaluation system. Faculty evaluators were a fixed group of senior staff physicians. Additionally, most faculty evaluations were created as a group evaluation wherein a representative faculty member from each service assimilates input from the other faculty members of the service into a single unified group evaluation. The scores for each question were combined to create an individual competency mean. The competency means were averaged and compiled. Individual question scores were also compiled. In addition, combined means of MK and PC were created, as well as for PR and ICS.

We attempted to validate our evaluation process by inviting program directors from outside our department to review the questions and behavioral prompts, which allowed for establishing face validity of our questions. We then further validated the content of the questions by having the evaluators read through the questions and determine the importance of that question in evaluating the residents, and then we used the consensus to include or exclude questions as described by Lawshe.16

A dataset including resident demographics and performance metrics was created. A separate dataset including a single academic year’s evaluations was also constructed. To evaluate the original hypothesis, correlations were used to explore the relationships between mean competency scores. Linear regression was used to establish the predictive power of PR and ICS on MK and PC ratings. In addition to faculty evaluation scores, aggregate scores from peer evaluations and nonphysician group evaluations were also used to compare to the ICS and PR competencies (Table 1). Data were compiled in Excel 2003 (M-
RESULTS

There were 37 residents for whom evaluations were available. We first examined whether there were any trends in how residents were evaluated based on their year of postgraduate training. It would not be unexpected to see an upward trend as residents progress through a training program and become more proficient in the necessary skills. Although no correlation was found between postgraduate year (PGY) and scores on MK, PR, PC, or ICS for faculty evaluations, a significant correlation was found between postgraduate year (PGY) and scores on MK, PR, or ICS for peer evaluations. This finding suggests that residents act differently when faculty are present. Additionally, PR score were evaluated as a proxy for resident affability. Briefly, the description provided by the ACGME for the ICS competency is that residents must demonstrate skills that result in effective information exchange and teamwork.13 We found a significant relationship existed between ICS and PR scores and a composite score of clinical performance (p < 0.001). To identify which factors were the most important in determining overall resident performance, we then created a model using individual questions in the predictive side with the dependant variable being a mean combined score of MK and PC. Linear regression revealed that PR and ICS scores positively influenced MK and PC scores (p < 0.001). In that model, 4 questions emerged as significant (individual p-values in Table 4). Another regression model was constructed using only these 4 questions that was also significant (p < 0.001). (Summary of Regression Results in Table 4). Overall, the model was significantly predictive and suggested an influence of softer competencies on the more clinical competencies.

Interestingly 1 question had a negative beta in the larger model (Beta = −0.212, p = 0.011) and a minimally nonsignificant beta (Beta = 0.041, p = 0.358) in the focused model. This finding suggests a negative relationship on this question, where lower performing residents were actually scored higher. Such a trend raises the possibility that evaluations may be biased in either direction by the residents’ performance in the “softer” competencies (PR and ICS).

Having found a correlation between scores on ICS and PR and higher overall evaluation scores, we next studied whether these scores might also correlate with higher scores on the American Board of Surgery In-Training Exam (ABSITE). The ABSITE is designed to measure a resident’s general knowledge level of the fundamentals of basic science and clinical management of surgical disorders.17 Interestingly, we found no correlation between ABSITE scores and ICS or PR composite scores. Furthermore, no correlation was found between MK or PC scores and ABSITE scores. Using factor analysis, we next looked to see whether any individual evaluation questions could predict performance on the ABSITE. No individual question or group of questions predictive of ABSITE performance could be found. This result suggests that our evaluation process may be measuring aspects of resident performance different from those measured by the ABSITE. Alternatively, if we believe that the ABSITE accurately measure true

### TABLE 1. Competency Distributions of Standard Resident Evaluations

<table>
<thead>
<tr>
<th>Evaluation</th>
<th>Total</th>
<th>MK</th>
<th>PC</th>
<th>PR</th>
<th>ICS</th>
<th>PBL</th>
<th>SBP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faculty of resident</td>
<td>22</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Junior resident of senior resident</td>
<td>9</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Senior resident of junior resident</td>
<td>11</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

These numbers represent the total number of questions per competency that are included on the evaluation forms.

We first examined whether there were any trends in how residents were evaluated based on their year of postgraduate training. It would not be unexpected to see an upward trend as residents progress through a training program and become more proficient in the necessary skills. Although no correlation was found between postgraduate year (PGY) and scores on MK, PR, PC, or ICS for faculty evaluations, a significant correlation was found between postgraduate year (PGY) and scores on MK, PR, or ICS for peer evaluations. This finding suggests that residents act differently when faculty are present. Additionally, PR score were evaluated as a proxy for resident affability. Briefly, the description provided by the ACGME for the ICS competency is that residents must demonstrate skills that result in effective information exchange and teamwork.13 We found a significant relationship existed between ICS and PR scores and a composite score of clinical performance (p < 0.001). To identify which factors were the most important in determining overall resident performance, we then created a model using individual questions in the predictive side with the dependant variable being a mean combined score of MK and PC. Linear regression revealed that PR and ICS scores positively influenced MK and PC scores (p < 0.001). In that model, 4 questions emerged as significant (individual p-values in Table 4). Another regression model was constructed using only these 4 questions that was also significant (p < 0.001). (Summary of Regression Results in Table 4). Overall, the model was significantly predictive and suggested an influence of softer competencies on the more clinical competencies.

### TABLE 2. Correlation Between PGY and Evaluation Scores

<table>
<thead>
<tr>
<th>Correlation and Statistical Significance vs PGY</th>
<th>MK</th>
<th>PR</th>
<th>PC</th>
<th>ICS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Staff evaluations</td>
<td>0.18</td>
<td>0.051</td>
<td>0.20</td>
<td>0.17</td>
</tr>
<tr>
<td>Peer evaluations</td>
<td>0.53</td>
<td>0.14</td>
<td>0.42</td>
<td>0.41</td>
</tr>
<tr>
<td>NS, not significant</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
</tbody>
</table>

NS, not significant.

There was no correlation between higher PGY status and evaluation scores among the staff evaluators. However, in the peer evaluation group, there was a significant correlation between PGY status and MK, ICS, and PC scores.
The high degree of intercorrelation among the 4 competencies studied—MK, PC, PR, and ICS—suggest that faculty could not effectively differentiate among these competencies when rating resident performance. These results are consistent with a substantial body of research on rater errors, including “halo error” (when salient aspects of a resident’s performance on 1 competency bias the evaluator’s judgments of the resident’s performance on other, unrelated competencies), “central tendency error” (the tendency to rate all residents as being “average” performers), and the “recency effect” (when recent examples of the resident’s behavior biases the evaluator’s recall of the resident’s earlier behaviors).18 The effects of rater errors have been known for over 60 years, when assessment centers were first used by the Office of Strategic Services (OSS) to select candidates for clandestine missions in occupied Europe.19 Similar findings have been observed in other domains, such as assessing pilot crew performance in commercial aviation20,21 and assessing management potential in the business community.22,23 Only recently, however, have practical recommendations been codified for developing appropriate observer-based performance measures and training raters to avoid these common rater biases. Measurement development requires multiple steps. Initially a detailed task analysis is performed to identify the critical skills that are to be measured. Then a series of Subject Matter Expert workshops are pursued to develop behavioral exemplars for each competency that are easy to observe and conceptually distinct. The necessary evaluation forms to minimize the rater’s level of cognitive workload are then designed. Finally, the measures are iteratively pilot tested/revised until they demonstrate psychometrically acceptable levels of reliability and validity.20,24 With regard to rater training, it will often

<table>
<thead>
<tr>
<th>Question Number</th>
<th>Text in Question</th>
<th>Beta</th>
<th>Beta Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRS3</td>
<td>The resident demonstrates ownership, responsibility, and personal accountability.</td>
<td>0.293</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>PRS5</td>
<td>The resident has demonstrated honesty, integrity, character, and trustworthiness during this rotation.</td>
<td>0.169</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>ICS4</td>
<td>The resident accepts and acts on constructive criticism.</td>
<td>0.260</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>ICS1</td>
<td>The resident demonstrates caring/empathy for patients.</td>
<td>−0.041</td>
<td>0.358</td>
</tr>
</tbody>
</table>

There was a significant correlation globally between higher scores on individual competencies and residents’ evaluation scores. Correlation is significant at the 0.01 level (2-tailed). Then linear regression was used to show predictive capabilities of the relationship and the model.
be necessary to implement behavioral observation training (BOT) to ensure that faculty evaluators can accurately detect, perceive, recall, and recognize specific examples of the residents’ behavior.\textsuperscript{25} Frame of reference (FOR) training is also required to provide faculty evaluators with a common set of standards for evaluating the residents’ performance.\textsuperscript{26}

This study was designed to look for correlations between scores in the PR and ICS competencies and overall resident performance. The hypothesis was that residents who were better able to communicate and participate on a team (affability) and who were more diligent in completing assigned tasks and making themselves available (availability) would score higher overall on their global evaluations. The study did demonstrate that PR and ICS scores correlated with higher scores in the other competencies; however, the underlying reason is unclear. It was difficult to ascertain whether residents with better ICS scores had higher PR, PC, and MK because of the halo effect, improper completion of evaluations, or whether those residents were truly performing better clinically. Indeed it is possible that a higher score on any of these 4 competencies may serve as a “halo” to draw a resident’s score upward in the others. In an attempt to clarify the situation, we examined scores from other evaluation sources. Data from peer evaluations correlated with staff evaluations except for ICS. This discrepancy may be from the residents’ changing behaviors in the presence of staff. Investigation of 360\degree evaluations, where both patients and ancillary staff participate in the process, may confirm whether this finding is the result of a deficiency in the evaluation process or of residents changing behavior patterns based on their interactions. Additionally, peer evaluations, in our study, seem to be biased based on PGY status, which may limit their usefulness in globally evaluating performance. Finally it must be pointed out that although we have attempted to validate our evaluations, further studies of inter-rater reliability and validity by comparisons to outside measures need to be performed. Perhaps as the data from national databases such as NSQIP become more available, we can compare graduate performance to how they were truly performing better clinically. Indeed it is possible that the less than rigorous feedback provided), or whether it wants an improved evaluative process with better quality measures and more helpful feedback. The results of this decision can have a profound impact on how we assess our residents’ performance, and by extension on the quality of care they provide their patients. Few programs have assembled the resources, commitment, and drive to make this leap. Moving away from educational volunteerism and borrowed educational time will be necessary and very expensive. The choice is ours to make, and the right answer is clear, albeit somewhat costly. The medical education community must have the courage to make the right decision.

\textbf{REFERENCES}

1. Lynch DC, Swing SR. Accreditation Council for Graduate Medical Education. Outcome Project Key Considerations for Selecting Assessment Instruments. Available at: http://www.acgme.org/outcome/assess/keyConsider.asp.

2. Likert R. A technique for the measurement of attitudes. \textit{Arch Psych}. 1932;140:5-53.


