North Pacific Surgical Association

Self-evaluation: how well do surgery residents judge performance on a rotation?

Kenneth W. Gow, M.D.*

Division of Pediatric General and Thoracic Surgery, Seattle Children’s Hospital and the University of Washington, 4800 Sand Point Way NE, Seattle, WA 98105

KEYWORDS:
General surgery; Resident; Education; Evaluation

Abstract

BACKGROUND: Surgical trainees are evaluated based on the Accreditation Council for Graduate Medical Education 6 core competencies. The ability for a learner to recognize strengths and weaknesses in these areas allows for critical self-improvement.

METHODS: Surgery residents rotating on a pediatric surgery rotation for 1 academic year were asked at an exit interview to provide a self-evaluation within the 6 core competencies on a Likert scale from 1 to 5. Self-evaluation scores were compared with a final group consensus attending evaluation. Further analyses included comparing residents as follows: less than R3 (junior residents) versus R3 (senior residents) residents, general surgery versus non–general surgery residents, university versus community residents, residents in the first half of the academic year versus residents in the second half, and top one third– and lowest one third–performing residents. Statistical analysis was performed using Student t tests with significance at $P < .05$.

RESULTS: A total of 45 surgical residents (29 junior residents and 16 senior residents) gave overall self-evaluation scores that were lower than attending evaluations (3.4 vs 3.8, $P = .0002$). This underscoring occurred for most core competencies, especially medical knowledge, operative skills, and practice-based learning but not professionalism or communication. When sorting residents by variables, there was significant underscoring by senior residents, general surgery residents, and highest one third–performing residents compared with junior residents, non–general surgery residents, and lowest one third–performing residents. There were no differences between self-evaluations and attending evaluations when comparing university with community residents and residents in the first half of the academic year with residents in the second half of the academic year.

CONCLUSIONS: Residents appear to have a more critical self-analysis than attending surgeons, with senior residents, general surgery residents, and highest one third–performing residents being the most critical of their own performance. Poorly performing residents appeared to lack insight into their abilities. This method of self-evaluation helps trainees reflect on their performance and highlights trainees who lack self-awareness and need counseling for improvement.

© 2013 Elsevier Inc. All rights reserved.
cycle to make final judgments about the learner’s success in achieving the objectives. The stakes involved in formative assessment are low because this assessment is meant to help the learner improve. However, the stakes in summative assessment are high, leading to a decision about success or failure in the learning unit.

It is the expectation of the Accreditation Council for Graduate Medical Education (ACGME) that the trainee will receive a written evaluation documenting areas of strengths and weaknesses noted during the training period. The evaluation will summarize the progress and incorporate feedback provided by educators on the services. The evaluation process is meant to foster learning and provide the trainee an idea of where he/she is in his/her development toward competency. All learners are expected to participate in lifelong learning and professional development. However, to develop, one must be able to identify strengths and weaknesses in knowledge, attitudes, and practice. Although a 360-degree evaluation provides a thorough evaluation of a trainee, the process does not include the trainee’s viewpoint. If the trainee is able to reflect into his/her own areas of weakness, he/she may work toward improvement. Conversely, if he/she does not recognize areas of weakness, then a trainee may be less likely to work on these areas. The goal of this project was to determine how well trainee self-evaluation (SE) correlated with faculty assessment.

Methods

Residents rotating on the pediatric general and thoracic surgery rotation at Seattle Children’s Hospital, Seattle, WA, during the academic year 2011 to 2012 were included in this study. Residents who rotate onto this service include residents ranging from R1 through R3, academic- and community-based residencies, and those comprising general surgery– and non–general surgery–type residents. The service includes 2 ACGME pediatric surgery fellows as part of the team, but they were not included in this study. There are 11 pediatric surgeons who are clinically active, and each works with the residents in an equal fashion in the operating room, clinic, emergency department, and wards. Each surgeon has input on the performance of each resident, and in each year 5 are selected to enter the online summative evaluation. An evaluation form was developed that matched the evaluation items found on the online resident management system (MedHub, MedHub Inc, Ann Arbor, MI) (Fig. 1). The evaluation form is a Likert scale from 1 through 5, with higher scores representing better performance. Categories included were those that assessed the 6 core competencies: patient care (clinical and operative), medical knowledge, practice-based learning and improvement (PBLI), professionalism, interpersonal skills and communication, and systems-based practice. Because patient care comprises 2 distinct areas of assessment (ie, clinical and technical), there were 7 evaluation categories. Information from weekly reviews of trainees was used along with comments from other personnel who interacted with the trainee (eg, nursing staff, operating staff, fellows, and attending surgeons) to develop a consensus 360-degree evaluation of the trainee (attending evaluation [AE]). During each trainee’s 1-on-1 exit interview, the trainee was asked to fill out an (SE). After the form was completed, a discussion ensued based on the SE and the AE. More time was spent on areas that either the trainee or the staff identified as areas needing improvement. A different attending surgeon from the one who performed the exit interview submitted the final evaluation form. The SE and the AE values were then tabulated into an Excel (Microsoft, Redmond, WA) spreadsheet for analysis. Further variables were included to determine whether there were any differences based on any resident variables (ie, level of resident, university-based residents [URs] or community-based residents [CRs], first or second half of academic year, general surgery [GS] or non–general surgery [NG] residents, and whether residents were low or high performing). Because the number of residents was too small to separate out as distinct types, categoric and preliminary general surgery R1 trainees were combined into the general surgery group.

Statistical analysis

Comparisons for the entire group included comparing the average for each category between the trainees’ SEs with the trainers’ AEs. The data also were resorted into the different resident variables. Comparisons for these resorted data were 3-fold: SE between the 2 different resident types, AE between the 2 different resident types, and the differences between the SE and AE between each resident type. Statistical analysis was performed using the Student t test to compare averages between groups with statistical significance set at \( P < .05 \).

Results

Overall

A total of 45 surgical residents (29 junior residents [JRs] and 16 senior residents [SRs]) gave overall lower SE scores than AE scores (3.4 vs 3.8, \( P = .0002 \)). This underscoring occurred for most core competencies including patient care–clinical, operative skills, medical knowledge, PBLI, and systems-based practice but not professionalism or communication (Table 1).

Junior residents or Senior residents

There were 29 residents in the R1 and R2 levels (JRs), and there were 16 R3 (SRs). JRs and SRs gave similar SE scores in all of the 7 categories. When looking at the AEs, SRs received higher scores in professionalism compared with JRs. When comparing SEs with AEs, SRs significantly underscored their SEs compared with JRs in 2 specific areas: professionalism and PBLI.
University-based residents or Community-based residents

There were 37 URs compared with 8 CRs. There were no significant differences when comparing UR with CR SEs in any of the categories. When looking at the AEs, there was no difference in scores between URs and CRs. When comparing SEs with AEs, there were no statistical differences among the URs and CRs.

First half or second half of academic year residents

There were 23 residents who began their rotation in the first half of the academic year and 22 in the second half of the academic year. There were no significant differences when comparing first with second half of the academic year resident SEs in any of the categories. When looking at the AEs, there was no difference in scores between residents in the first and second half of the academic year. When comparing SEs with AEs, there were no statistical differences among residents in the first and second half of the academic year.

General surgery or Non–general surgery residents

There were 23 GS residents and 22 NG residents who rotated during the academic year. Regarding SEs, there were no statistical differences in SE scores, with the only exception being NG residents scoring themselves lower on

---

**Figure 1** The form that residents were asked to perform self evaluation.
medical knowledge compared with GS residents. Overall, GS residents received higher AE scores in all areas compared with non-GS residents, with only medical knowledge being not statistically significantly different. When comparing the AE scores with SE scores, GS residents scored themselves lower than NG residents in all areas with the exception of patient care–clinical and medical knowledge.

**Low- versus high-performing residents**

The residents were sorted based on the average AE scores into upper, middle, and lower thirds. The residents within the highest third (HI) were compared with the lowest third (LO). With regards to SEs, there were no significant differences between HI- and LO-performing residents. With regards to AEs, the higher-performing residents were higher in all categories compared with the lowest-performing residents. When comparing AEs with SEs, LO residents scored themselves higher than the AE in professionalism, interpersonal skills and communication, and systems-based practice (Table 2), whereas HI residents scored themselves lower in all categories compared with their AEs (Table 3).

**Comments**

The surgical trainee is evaluated in several different areas during his/her training. The summative evaluation at the end of a rotation forms one of the means by which a trainee will know how he/she has done and directs the learner to focus on areas for improvement. However, another area that has also been used by some educators is the concept of reflection as a means for residents to consider their own strengths and weaknesses. Schon first championed the use of reflection in medical education. Reflection allows for a deep level of understanding and may help channel more meaningful change. By having a resident evaluate himself/herself, there was an attempt to merge these 2 concepts into 1 meeting to provide maximal benefit for the trainee and the educator.

Residents gave themselves lower scores than the evaluating attending physicians, which align with other similar reviews of self-assessment. In a meta-analysis of 44 self-assessment studies in higher education, there was a .39 correlation between self-assessment and expert assessment, with participants generally underrating their performance. Some possible explanations for this mismatch in scores include the following:

1. Trainee-related reasons including limited opportunities for true comparisons with other trainees at their level, which may limit their ability to benchmark; an inner drive for improvement, which may be interpreted as being self-critical; or a demonstration of modesty.
2. Attending-related reasons including expectations not sufficiently spelled out leading to trainees assuming a higher level of performance than was expected.
3. Process-related reasons including the setting of SE an exit interview may have been one in which the trainee believed he/she was required to be self-critical.

These explanations highlight the need to set reasonable expectations. Although this may be difficult to do in a microenvironment, giving residents reasonable expectations would be a good means to reduce the mismatch seen here.

The area in which mismatch was most intriguing was in the HI- and LO-performing groups. For the top performers, the trainees gave themselves lower marks than the attending did. This may imply that high performers are setting high standards, they are hypercritical, or they may overlook their own performance. This is in distinction to the lower

---

**Table 1** Overall differences between SEs and AEs

<table>
<thead>
<tr>
<th></th>
<th>Patient care–clinical</th>
<th>Operative skills</th>
<th>Medical knowledge</th>
<th>PBLI</th>
<th>Professionalism</th>
<th>Interpersonal skills and communication</th>
<th>Systems-based practice</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean SE</td>
<td>3.4</td>
<td>3.0</td>
<td>2.9</td>
<td>3.4</td>
<td>3.8</td>
<td>3.8</td>
<td>3.2</td>
<td>3.4</td>
</tr>
<tr>
<td>Mean AE</td>
<td>4.0</td>
<td>3.6</td>
<td>3.8</td>
<td>4.0</td>
<td>3.9</td>
<td>3.9</td>
<td>3.6</td>
<td>3.8</td>
</tr>
<tr>
<td>Statistical analysis</td>
<td>&lt;.0001*</td>
<td>&lt;.0001*</td>
<td>&lt;.0001*</td>
<td>&lt;.0001*</td>
<td>.6158</td>
<td>.5018</td>
<td>.0082*</td>
<td>.0002*</td>
</tr>
</tbody>
</table>

*P < .05 between SEs and AEs.

**Table 2** Differences between SEs and AEs in the LO-performing residents

<table>
<thead>
<tr>
<th></th>
<th>Patient care–clinical</th>
<th>Operative skills</th>
<th>Medical knowledge</th>
<th>PBLI</th>
<th>Professionalism</th>
<th>Interpersonal skills and communication</th>
<th>Systems-based practice</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean SE</td>
<td>3.4</td>
<td>2.9</td>
<td>2.8</td>
<td>3.5</td>
<td>3.7</td>
<td>3.8</td>
<td>3.2</td>
<td>3.3</td>
</tr>
<tr>
<td>Mean AE</td>
<td>3.1</td>
<td>2.8</td>
<td>3.1</td>
<td>3.1</td>
<td>3.1</td>
<td>3.0</td>
<td>2.9</td>
<td>3.0</td>
</tr>
<tr>
<td>Statistical analysis</td>
<td>.1632</td>
<td>.6183</td>
<td>.0631</td>
<td>.0911</td>
<td>.0053*</td>
<td>.0033*</td>
<td>.0247*</td>
<td>.0150*</td>
</tr>
</tbody>
</table>

*P < .05 between SEs and AEs.
performers who scored their performance higher than the attending did. These paradoxical findings are similar to another study in which GS residents were asked to provide a self-assessment.\textsuperscript{13} This is concerning because if a trainee does not think he/she has anything to improve, it is unlikely that improvement will occur, but educators may use this information to provide the lowest-performing residents the help they need to improve. Program directors need to be cognizant that residents performing at the lowest levels may be the hardest to improve. Until residents have reflected on and assimilated the feedback, plans for learning and change are not likely to materialize.\textsuperscript{13} Further hampering this process is that physicians respond with negative emotions to feedback that they feel is inconsistent with their self-perceptions of performance or that questions their credibility, and, therefore, they are not inclined to apply it in a constructive manner.\textsuperscript{14}

The other areas in which there were mismatches were GS versus NG residents and SRs versus JRs. In both situations, the SE scores were similar, whereas the AE scores were higher with the GS residents and SRs. This may point out some biases by the trainees and evaluators. First, trainees, when faced with SEs, may automatically place themselves as average, which is known as “central tendencies.”\textsuperscript{15} As for the attendings, there may be another effect that leads to skewing GS residents and SRs above their NG and JR counterparts. There may be a “better than average” phenomenon (also known as the “Lake Wobegon effect”) in which evaluators perceive that all GS residents and especially SRs who are all within GS training programs are “above average.”\textsuperscript{12} Also factored into this is the “halo effect” in which residents with reputations are scored based on those expectations as opposed to true performance.\textsuperscript{15} However, because this effect may skew results both higher (good reputation) and lower (poor reputation), this effect may be minimal.

This technique of reflection may be useful for final evaluations for both the evaluator and the trainee. For the evaluator, it may provide an excellent set point for discussion on the trainee’s performance during a rotation. This may produce more efficient and effective discussions because one may then specifically focus on areas in which the trainee’s assessment does not match the attending’s assessment. For the trainee, the role of reflection may be either introduced or encouraged.\textsuperscript{16} These techniques are important for self-improvement both during training and for later when they are in practice.\textsuperscript{17}

The format of evaluation was based on the 6 core competencies and was easily understood by evaluators and trainees alike. The ACGME Milestone Project is an evolution toward an outcome-based method for evaluating resident performance.\textsuperscript{18} It defines the specific behaviors, attributes, and outcomes within the 6 general competency domains to be demonstrated by residents at particular points during their education; these sets of expectations are known as “milestones.” Although this may be a better method for yearly evaluations, trying to make this applicable to rotational evaluations may be challenging: residents are used to defining how they are performing based on a simple scale and will need to grow accustomed to a more global view. Regardless of these changes, self-reflection on performance can still be used assuming both trainees and educators can grow accustomed to the new evaluation methodology.

### Limitations

This is a small study performed with 1 year of residents rotating on 1 rotation in 1 training program. Therefore, trying to generalize these results to other residents and other programs would be unwise until further studies are performed. However, this technique will be continued to see if the findings from this study are replicated in different cohorts of trainees. Furthermore, although the number of attending surgeons entering the resident evaluation scores was kept to a minimum, having different evaluators does introduce inter-rater variability, which may affect the results.\textsuperscript{19} It is well known that despite an attempt to make evaluations as objective as possible, it is hard to argue that most continue to include a significant component of subjectivity.\textsuperscript{20}

### Conclusions

Overall, residents appear to have more critical self-assessments compared with how attending surgeons assess the residents. Areas in which this was highlighted included GS, senior, and high-performing residents being the most critical of their own performance. Conversely, poorly performing residents overscored themselves and appear to lack insight into their abilities. This method of SE is useful because it highlights those trainees whose lack of self-awareness will perhaps require counseling in order to improve.
References


Discussion

Mika Sinanan, M.D., Ph.D. (Seattle, WA): The provision of feedback to trainees in all disciplines is an important element in helping them to progress, to gain competency, and to eventually approach mastery of their chosen field. In medicine and perhaps especially in surgery, it seems that we do not do a particularly good job of providing actionable feedback to our trainees, at least in a format and setting where they perceive it as constructive and supportive. In this study, a group of residents rotating on a pediatric surgical service over a 1-year period had their own perceptions of performance (self-assessment) compared with a “360 assessment” compiled with input from faculty and other health care team members and delivered after the service rotation was completed. The 360 assessment was documented through a survey tool using the same format and categories as the self-assessment based on defined ACGME competencies. Although this study is limited by the small number of residents and faculty (all from a single institution over a single year), several interesting observations deserve to be highlighted and may be worthy of further investigation. First, residents as a group tended to rate their own performance lower or more conservatively than others (faculty and nonphysician health care team members) who objectively observed their performance. Second, SRs personally rated their performance lower when compared with faculty and staff than JRs in certain categories, specifically professionalism and practice-based learning. Third, residents stratified into high, medium, or low performers based on the faculty and staff evaluations had some important differences in perception of their own performance. High performers rated themselves lower in all categories compared with the faculty and staff, whereas low performers rated themselves more highly than the faculty and staff in 3 categories: professionalism, interpersonal skills and communication, and systems-based practice. The assessment literature highlights the common finding that self-assessment often delivers a lower rating of performance than objective assessment. Reasons for this finding remain unclear but in surgical disciplines they probably include our perfectionistic culture and orientation that selects goal-directed individuals who might reflexively be harsher judges of their own performance. Recent leadership training studies suggest that individuals assessing their own performance tend to compare themselves with an ideal, whereas others are more objective, comparing them not with an intention but rather with observed behavior, often with the last group of trainees or peers. These 2 vastly differently standards of comparison might account in part for the observed global differences in the current study and the apparent amplifying effect of becoming a more SR on self-critical assessment. Other differences in the 2 types of summative assessment might also bear on this analysis. Summative evaluations by trainers are episodic, whereas SE and criticism tend to be continuous although they were actually measured only at 1 time point in this study. Self-assessment likely includes more of an emotional component. Comparison with an expected level of performance can quickly become a comparison to the recollection of prior performance that introduces standards that differ between individual trainees. These are all issues that need to be considered in using assessment tools to advance the goals of a training program for mission-critical skills. I have 3 questions for Dr. Gow. First, how do you plan to further validate the assessment tools and confirm these early observations? Second, based on your study, how might self-assessment and objective assessment both offer value and be used to serve constructive roles in a training environment? Third, does the apparently inflated perception of skill for low performers in certain domains suggest any ways we should consider modifying our surgical training programs?