Feasibility of an Internet-Based Global Ranking Instrument

Seshadri C. Mudumbai, MD
David M. Gaba, MD
John Boulet, PhD
Steven K. Howard, MD
M. Frances Davies, PhD

Abstract

Background Single-item global ratings are commonly used at the end of undergraduate clerkships and residency rotations to measure specific competencies and/or to compare the performances of individuals against their peers. We hypothesized that an Internet-based instrument would be feasible to adequately distinguish high- and low-ability residents.

Materials and Methods After receiving Institutional Review Board approval, we developed an Internet-based global ranking instrument to rank 42 third-year residents (21 in 2008 and 21 in 2009) in a major university teaching hospital's department of anesthesiology. Evaluators were anesthesia attendings and nonphysicians in 3 tertiary-referral hospitals. Evaluators were asked this ranking question: "When it comes to overall clinical ability, how does this individual compare to all their peers?"

Results For 2008, 111 evaluators completed the ranking exercise; for 2009, 79 completed it. Residents were rank-ordered using the median of evaluator categorizations and the frequency of ratings per assigned relative performance quintile. Across evaluator groups and study years, the summary evaluation data consistently distinguished the top and bottom resident cohorts.

Discussion An Internet-based instrument, using a single-item global ranking, demonstrated feasibility and can be used to differentiate top- and bottom-performing cohorts. Although ranking individuals yields norm-referenced measures of ability, successfully identifying poorly performing residents using online technologies is efficient and will be useful in developing and administering targeted evaluation and remediation programs.

Introduction

Global rating instruments (GRIs) are commonly used at the end of undergraduate clerkships or residency rotations to assess overall competence.[1,2] Widely accepted and easy to use, GRIs assess interpersonal and communication skills, professionalism, and aspects of patient and systems-based care.[3–6] GRIs may be relevant in dynamic domains like anesthesiology, where standardized educational tools such as written and oral exams or objective, structured, clinical examinations may not directly measure important aspects of clinician performance.[7,8]

For example, consider an anesthesiology resident faced with unexpected, massive hemorrhage in the operating room (OR). Successful management skills include simultaneously coordinating and communicating care strategies with surgical and nursing teams, performing multiple invasive procedures, and working under time pressures. A multiple-choice test can probe static knowledge about hemorrhage, an oral examination can measure choice of management strategies or clinical decision making, and an objective, structured, clinical examination can measure the performance of procedures. However, none of these strategies probes the integration and dynamism of the work, and evaluations may not correlate with a clinician's actual job performance.[9–11] Thus, a GRI in which multiple evaluators assess anesthesia residents' job performance—delivery of care throughout the perioperative period and overall clinical ability, including nonprocedural skills such as communication and team leadership[12]—may be valuable.

Despite these potential benefits, global ratings present difficulties: First depending on evaluators' training and dedication, ratings (scores) can be too homogeneous to discriminate between individuals.[1] Second, because remediating poor performers is time-consuming, evaluators may tend to rate all individuals near the top of the scale.[1,3]

Seshadri C. Mudumbai, MD, is Staff Anesthesiologist at VA Palo Alto Health Care System and Instructor of Anesthesiology at Stanford University; **David M. Gaba, MD,** is Staff Anesthesiologist and Director of Patient Simulation Center of Innovation at VA Palo Alto Health Care System and Associate Dean for Immersive & Simulation-based Learning and Professor of Anesthesia at Stanford University; **John Boulet, PhD,** is Associate Vice President for Research and Data Resources at the Foundation for Advancement of International Medical Education and Research; **Steven K. Howard, MD,** is Staff Physician at VA Palo Alto Health Care System and Associate Professor of Anesthesia at Stanford University School of Medicine; and **M. Frances Davies, PhD,** is Research Associate Director of Faculty Development at VA Palo Alto Health Care System.

Corresponding author: Seshadri C. Mudumbai, MD, Stanford University/VA Palo Alto HCS Anesthesiology, 3801 Miranda Av(112 A), Palo Alto, CA 94304-9891, mudumbai@stanford.edu

Received August 24, 2010; revision received October 10, 2010; accepted October 26, 2010.

DOI: 10.4300/JGME-D-10-00162.1

Third, when GRIs contain numerous items measuring several competencies, evaluators may not complete all of the evaluations.[13] Finally, if asked to rate many subjects, especially using paper-based systems, evaluators may get confused unless properly queued and alerted as to who is being rated. Fortunately, given widespread Internet access,[14,15] online evaluation tools present opportunities to address these problems for GRI by taking advantage of computer applications that ensure proper completion, ease of delivery over the Web, and use of interactive text and images.

We hypothesized that such an Internet-based instrument that proceeded immediately to a single-item global ranking would be feasible for gathering data and identifying, albeit from a norm-referenced perspective, high- and low-performing anesthesiology residents. Because our intent was to evaluate the technical viability of this approach, our first goal was to collect a sufficient quantity of rankings from different evaluator groups. Then, we explored whether the rank-orderings were consistent and could be aggregated to discriminate individuals along the ability continuum.

Methods

We created an Internet-based, data-collection tool designed to obtain relative ability rankings of 2 consecutive 21-member classes of third-year anesthesiology residents at Stanford University Medical Center. Residents rotate among 3 hospitals (University, Children's, and Veterans Administration) during their 3-year residency. We implemented our ranking instrument during the last half of residents' third year (once in Spring 2008 and again in Spring 2009), after their completion of general and specialty rotations (general surgery, neurosurgery, obstetrics, pediatrics, cardiac surgery, intensive care medicine, and pain management). The 2 evaluator groups, who regularly worked in the OR environment, were anesthesiologist-physicians (''attendings'') and nonphysicians (''nonphysicians''). Nonphysicians were OR and recovery room nurses, scrub and anesthesia technicians, and OR nursing administrators. The project received approval by the Stanford University Institutional Review Board.

In both resident groups the average age was 32 years. Men outnumbered women (in 2008: men = 11 and women = 9; in 2009, men = 13 and women = 8). In the 2008 class, 3 residents had other graduate degrees (MSc or PhD), and 4 had completed another residency training program (eg, medicine, pediatrics). In the 2009 class, 5 residents had earned graduate degrees in some discipline, and none had completed another residency.

The Institutional Review Board approval included a waiver of consent for the resident cohorts in the study. This was important in avoiding selection bias by preventing residents from either opting in or out to be ranked. All residents could potentially be ranked. Although evaluators knew residents' identities, residents' confidentiality and rankings were protected (see later). Evaluators, who could anonymously choose not to participate, gave implied consent, by submitting their rankings.

Ranking Instrument Development

Global Ranking Question Before we decided on 1 or more global ranking questions or the technical aspects of an online application, we first surveyed our potential evaluators, who told us that addressing multiple questions would be too time-consuming. So we selected a single global ranking question: ''When it comes to overall clinical ability, how does this individual compare to all their peers?'' This question, which addresses clinical management skills exhibited during the entire perioperative period, is aimed at determining if rankings based on a single, holistic question could differentiate overall resident performance.

Our evaluators indicated it would be difficult to rank-order every cohort member along a continuum from lowest to highest, even if they referenced a single global construct. Moreover, the prevailing literature indicated a specific score (ie, rating) would risk leniency or severity bias (a tendency to score too easy or hard) or central-tendency bias (not using the entire scale).[1,3] Therefore, we instructed evaluators as follows: Identify residents who couldn't be ranked. Consider each resident's ability throughout the perioperative period—preoperative planning, intraoperative management and team leadership, and postoperative management. Then rank each resident as follows: Place each into 1 of 5 bins (quintiles) of relative class ranking: Quintile 1 (top of class, 81%–100%); Quintile 2 (61%–80%); Quintile 3 (middle of class, 41%–60%); Quintile 4 (21%–40%); Quintile 5 (bottom of class, 0%–20%). Each quintile must contain an equal (''balanced'') number of residents. If the remainder exceeds an even multiple of 5 (eg, 11, 17, or 18), place 1 remaining resident into a quintile (1 per quintile). For example, an evaluator who knew 18 residents would assign 3 per quintile (15 total) and then assign each of the 3 remaining ones to a different quintile.

Technical Development We then worked with Stanford's Information Resources and Technology to develop and host an Internet-based application to allow respondents to submit their evaluations securely and confidentially online. We selected the JAVA 4.2 programming language, which works in heterogeneous operating systems and browser environments (ie, Windows, MAC OS, Linux; Internet Explorer, Safari, and Mozilla Firefox).

The application was designed to maintain resident evaluation and ranking information as confidential. All communication was transmitted over a secure encrypted network using a secure socket layer in the web browser. Authentication was performed using the university's robust, web-authorization system. The application operated in a secure environment with the Oracle databases, Java application servers, and Apache web servers hosted on a

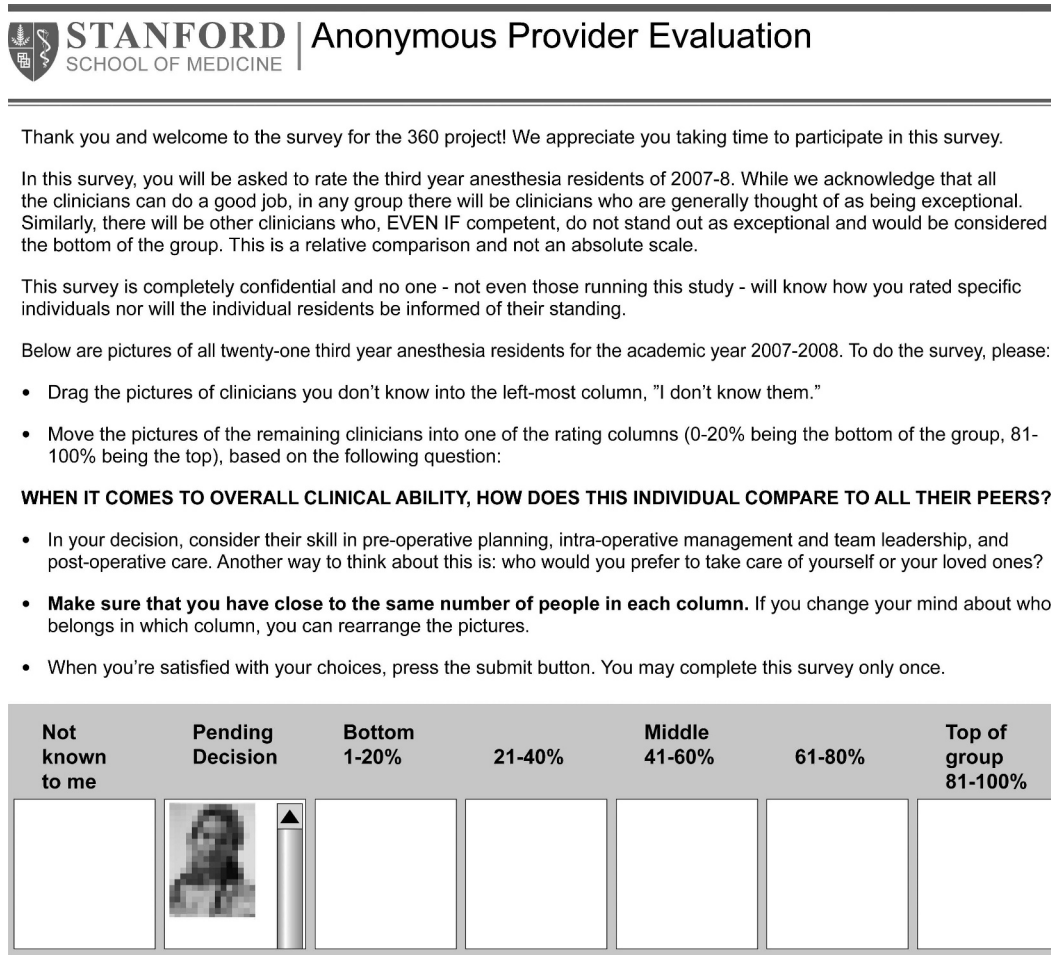

FIGURE 1 **Screenshot of Interface of Internet-Based Ranking Instrument**

The blurred square corresponds to the photo of a member of the resident cohort (not shown to preserve confidentiality). Once the photos were logged into the ranking instrument website, evaluators were presented pictures of the third-year anesthesia residents in a column below the text box. First, evaluators were asked to place residents they did not know into a column labeled "Not Known." Then, they were asked to drag and drop photos of the remaining residents into bins corresponding to quintiles.

secure network; network access was strictly controlled and monitored. Complete summary roll-up reports of resident rankings were delivered to investigators as deidentified Microsoft Excel spreadsheets.

To further ensure confidentiality, our online system did not capture identifying evaluator information that could be linked to a ranking. Rankings could not be changed, once submitted. Coded identifiers available to the research team delimited residents in the results data file. A neutral third party who knew neither residents nor evaluators was given access to the resident names and codes to develop a crosswalk file that linked ranking, demographic, and other performance data. Because of these protections, residents could never know who evaluated them, and investigators could not know the relative rankings of residents by name. Furthermore, evaluators could not know one another's rankings. The data were stored in a secure database, as if they were patient clinical trial data. In addition, because a crosswalk capability existed, we obtained a National Institutes of Health Certificate of Confidentiality to protect against compelled disclosure of identifiable rating information.

Pilot Testing and Survey Deployment For the pilot-testing phase, we developed and sent an explanatory e-mail about the study and the ranking instrument to a small pool of potential evaluators. The e-mail contained a hyperlink to the global ranking questionnaire. We also ensured that evaluators could access the questionnaire from computer workstations (in all 3 hospitals) in the OR, postanesthesia care unit, library, offices, and home computers. During this pilot phase, we learned that potential evaluators had difficulty identifying residents by name alone, so we provided a picture of each resident. FIGURE 1 depicts the instrument in its final version.

After pilot testing, we conducted presentations at the various services' departmental meetings to inform potential

evaluators and the resident cohort about the purpose of the project and to answer any questions.

Statistical Analysis

Using Microsoft Excel, we calculated overall response rate (number responded per potential number of respondents) and response rate for each evaluator group. For all available rankings for each resident, we calculated the following: arithmetic mean, mode, and median and measures of dispersion including variance, skew, kurtosis, and interquartile range. From the median or mean of evaluator rankings (both overall and by evaluator group), we derived a summary rank order of the resident cohort. For each resident, we computed the fraction of a given evaluator type (attendings and nonphysicians) that placed the resident into each quintile. Computations, including correlations, were performed with Microsoft Excel or SAS 9.2 (SAS Institute Inc, Cary, NC). Scatterplots and linear regressions between the rankings of the 2 evaluator groups were produced using GraphPad Prism 4.0 (GraphPad Software Inc., La Jolla, CA).

Results

For 2008, 111 evaluators (attendings = 41, nonphysicians = 70) ranked the residents. For 2009, 79 evaluators (attendings = 27, nonphysicians = 52) did so. The potential evaluators were attendings (N = 103) and nonphysicians (N = 288) for both years. Of the nonphysicians who responded for both years, 83% were OR and recovery room nurses, 3% were nursing administrators, 10% were scrub technicians, and 4% were anesthesia technicians.

FIGURE 2 shows that both evaluator groups clustered certain individuals at the top (''bubbled up'') or bottom (''sunk down''). Residents clustered at the top were most often ranked in the top 2 quintiles and received few or no rankings in the bottom 2. This skewed distribution is similar to, but opposite from, what was seen for those considered to be at the bottom. Residents who were considered to be neither top nor bottom but in the middle tended to be ranked at least once, by at least 1 evaluator, in each of the 5 quintiles. Overall, the attendings and nonphysician evaluators were consistent in ranking certain individuals at the top and bottom of their classes; residents in the middle were ranked along the entire continuum.

Table 1 and Table 2 summarize attendings and nonphysician evaluations for each resident cohort. Both 2008 and 2009 rankings showed that certain residents were considered to be at the top or bottom for both sets of evaluators. In addition, the summary measure of relative performance indicated that no residents were categorized by 1 evaluator group (eg, attendings) as top performers who were conversely categorized by the other group (nonphysicians) as bottom performers. The number of evaluators per resident for each evaluator group for both year 2008 and 2009 was fairly consistent and showed a low variance.

FIGURE 3 shows the relationship between attending and nonphysician summary rankings for the 2008 and 2009 cohorts. Spearman rank correlations between the summary values for the 2 types of evaluators were moderate to high for both 2008 ($r = 0.46$, n = 21, $P < .05$) and 2009 ($r = 0.69$, n = 21, $P < .05$), suggesting that attending and nonphysician ability categorizations were similar.

Discussion

In this single academic center study, we examined the feasibility of anesthesia attendings and OR personnel to rank residents in overall ability using an Internet-based data collection system. Given our GRI's JAVA 4.2 development platform, our Internet-based categorization tool could be accessed by multiple computer operating systems and web browsers. Evaluators were able to complete the ranking task in large numbers, despite the fact that their participation was voluntary. Although the overall response rate, in absolute percentage for both years and groups, was approximately 30%, the denominator (pool of potential evaluators who interact with the residents) was very large as would be expected in implementing a study across 3, tertiary, referral hospitals. Because the number of evaluators for each resident was relatively large, and the 2 rater groups provided similar rankings, it is likely that the relative ability estimates would be generalizable at least for the top- and bottom-performing residents.

In comparison with other studies examining Internet-based scoring instruments,[15,16] embedding a hyperlink to a URL within an e-mail allowed us to bypass some of the issues of difficult access caused by hospital firewalls. The use of pictures of residents with an easy drag-and-drop interface is novel and aided evaluators in their completion of the ranking instrument.

We also examined the ability of the summary measures derived from this instrument to differentiate between top- and bottom-performing residents. First, as expected, our study showed that each evaluator group's rankings were clustered and for both classes of third-year residents. If a large proportion of evaluators assigned individuals to the top 2 quintiles and rarely to the bottom 2, we could reasonably assume that these individuals were high-performing clinicians for their level of training. Similarly, if a large proportion of evaluators assigned individuals to the bottom 2 quintiles, with few assigning them to the top 2 quintiles, these individuals were likely to be low-performing clinicians for their level of training. If attending and nonphysician rankings had been purely random, we would have expected nearly all residents to be nearly equally assigned to the 5 quintiles, with nearly identical mean or median summary rankings and little clustering.

Despite the clustering of resident ratings, evaluators' lack of uniform consensus on who should be assigned to the top and bottom cohorts suggests that some evaluators may have referenced different skills in completing the ranking

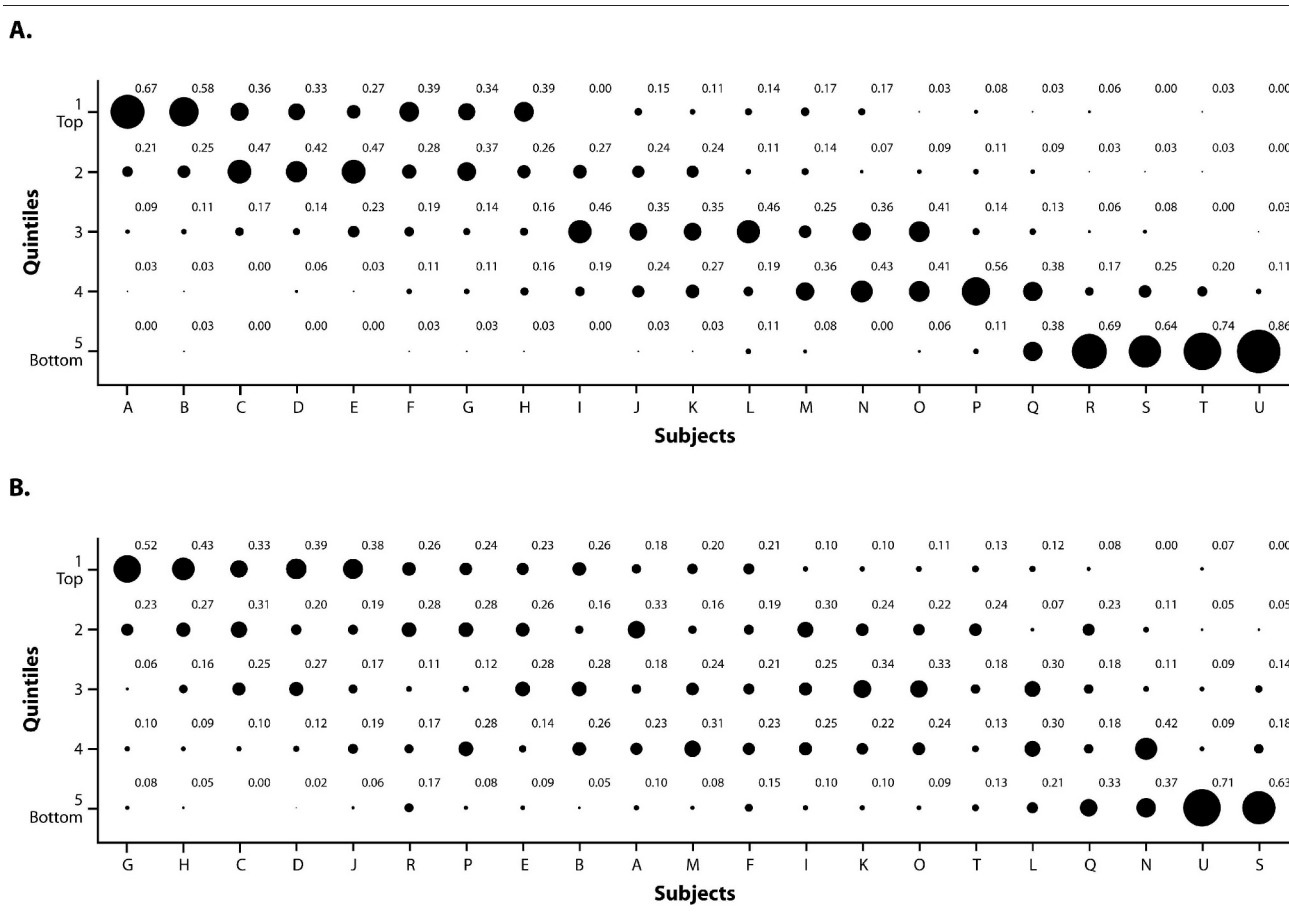

FIGURE 2A **Summary of Attending Ratings for Resident Class of 2008**

FIGURE 2B **Summary of Nonphysician Rankings for Resident Class of 2008**

Attendings were anesthesia attendings. The actual rankings, as frequencies that each resident received, are listed. Frequencies were the ratio of resident categorizations per quintile divided by total number of resident categorizations. Median scores were used to rank residents. Ties were broken by further ranking using ascending values of the arithmetic mean of rankings that the residents received. The diameters of the bubbles are proportional to the actual percentage of ratings per quintile that an individual received. Residents for anesthesia attendings are ranked as A through U. The x-axis showed the resident alphabetic code in increasing order of median score. The y-axis showed the frequency of rating in each quintile, represented by a dark circle whose radius is proportional to the frequency.

Nonphysicians were operating and recovery room nurses, scrub and anesthesia technicians, and operating room nursing administrators. The residents are the same residents that attendings rated for 2008. The actual rankings, as frequencies that each resident received, are listed. Frequencies were the ratio of resident categorizations per quintile divided by total number of resident categorizations. Median scores were used to rank residents. Ties were broken by further ranking using ascending values of the arithmetic mean of rankings that the residents received. The diameters of the bubbles are proportional to the actual percentage of ratings per quintile that an individual received. The x-axis showed the resident alphabetic code in increasing order of median score. The y-axis showed the frequency of rating in each quintile, represented by a dark circle whose radius is proportional to the frequency.

task. For example, nonphysicians may have ranked personality and communication skills higher than did attendings, whereas attendings may have ranked technical and procedural skills higher.[13,17] Another reason for a lack of uniform consensus may be that, although global ratings have higher reliability than behaviorally anchored ratings,[1] different evaluators may have varied in their ability to provide a summary judgment. Given that only a global performance measure was referenced in the ranking process, additional studies should be aimed at better understanding how evaluators make their judgments and what skill sets they favor in doing so. More important, from a construct validity perspective, quantifying the relationships between multisource rankings and actual clinical ability (eg, procedural skills, patient outcomes) are needed. Because our initial goal was to evaluate if high- and low-ability cohorts could be classified consistently, we did not take the step of identifying individuals who might benefit from additional educational interventions. This strategy of recognizing particular residents, which might also involve a more detailed rating of specific skills, could be used to better align remedial instruction or skills training with learners' needs.

Second, a key aspect of the rating process was the unique use of a binned, relative ranking based on a single global construct. We asked evaluators to place individuals

TABLE 1 **Rankings and Summary Statistics for the 2008 Surveys**[a]

Ratee	Attendings		Nonphysicians	
	Median Rating	Number of Raters per Resident	Median Rating	Number of Raters per Resident
A	1	33	3	40
B	1	36	3	43
C	2	36	2	48
D	2	36	2	51
E	2	30	3	43
F	2	36	3	47
G	2	35	1	48
H	2	31	2	44
I	3	37	3	40
J	3	34	2	47
K	3	37	3	41
L	3	37	4	43
M	3	36	3	49
N	3	14	4	19
O	3	31	3	45
P	4	36	2	50
Q	4	32	4	39
R	5	35	2	53
S	5	36	5	57
T	5	35	3	55
U	5	37	5	58
Summary Statistics	Mean = 34, SD = 5		Mean = 46, SD = 8	
	Minimum = 14; Maximum = 37		Minimum = 19; Maximum = 58	

[a] Absolute rankings that each evaluator group (attendings and nonphysicians) gave are listed; attendings are the reference group for the alphabetical coding of residents. Residents for 2008 are listed from A to U; residents for 2009 are listed from A1 to U1.

TABLE 2 **Rankings and Summary Statistics for the 2009 Surveys**[a]

Ratee	Attendings		Nonphysicians	
	Median Rating	Number of Raters per Resident	Median Rating	Number of Raters per Resident
A1	1	26	2.0	34
B1	1	27	2.0	35
C1	2	26	2.0	37
D1	2	27	1.0	42
E1	2	24	3.0	40
F1	2	24	3.0	37
G1	2	26	2.0	42
H1	3	25	3.0	26
I1	3	26	4.0	38
J1	3	24	3.0	40
K1	3	20	2.0	36
L1	3	27	2.0	43
M1	3	18	4.0	31
N1	3.5	26	2.5	44
O1	3.5	26	4.0	35
P1	4	24	3.0	35
Q1	4	25	4.0	40
R1	4	25	3.0	38
S1	4	25	3.0	43
T1	5	27	5.0	38
U1	5	27	4.0	40
Summary Statistics	Mean = 25, SD = 2		Mean = 38, SD = 4	
	Minimum = 18; Maximum = 27		Minimum = 26; Maximum = 44	

[a] Absolute rankings that each evaluator group (attendings and nonphysicians) gave are listed; attendings are the reference group for the alphabetical coding of residents. Residents for 2008 are listed from A to U; residents for 2009 are listed from A1 to U1.

into quintiles and so forced evaluators to rank residents relatively. Even though this approach is norm-referenced, we were able to identify individuals who were deemed to be much better, or worse, than their peers with respect to overall ability. By ranking residents, as opposed to rating them, we avoided leniency, severity, or central-tendency biases. Nevertheless, depending on the general competence of the cohort, and the choice of individuals who provide the rankings, it is still only possible to make general inferences regarding the competence of individual residents. However, in situations where training and remediation resources are limited, the ranking process is quick and efficient for identifying those individuals for whom additional evaluation and training may be needed.

Third, in formulating our global question, we intended to assess care delivered throughout the perioperative process and to test the utility of a single global question to rank the residents. Davis[18] found a global rating of obstetrics-

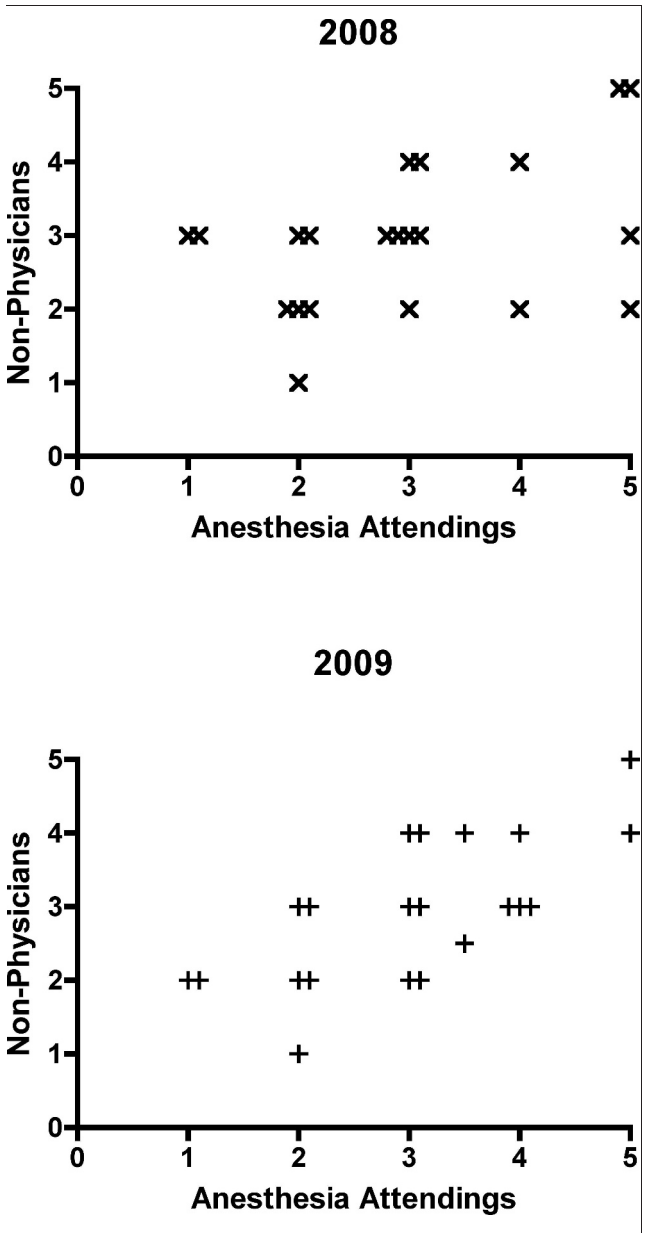

FIGURE 3 **Scatterplot of Attending and Nonphysician Summary Median Rankings per Resident for Instrument Administered in 2008 and 2009**

Nine residents received the same votes as another resident, resulting in only 12 unique points being shown.

gynecology residents that took into account both clinical competency and interpersonal skills was useful as an overall assessment rubric. In another study that examined residents rotating through an intensive care unit, researchers found high intra- and interclass correlations for attendings and peer residents but poor correlation between attendings and nurses for markers of overall competence.[19] However, given that global ratings have been listed by the Accreditation Council for Graduate Medical Education as part of its educational toolbox, we hope other researchers will consider using global questions in instrument development,[20–22] especially data-collection tools deliverable over the Internet. Given the logistical complexity of having evaluators rate, or rank, individual performances on multiple dimensions over time, it may be more prudent to limit the scope of the initial evaluation, especially if the top and bottom cohorts can be reliably identified. Where this can be accomplished, more detailed follow-up evaluations can be done.

Because our study demonstrated feasibility, it may be useful to consider where our GRI has a high, potential utility. This GRI could apply to many disciplines, including surgical and medical specialties (eg, orthopedics and invasive cardiology), which share with anesthesiology the challenges of integrating technical and nontechnical skills within dynamic environments. It could help rank residents earlier than their third year of residency, identifying weak cohorts to remediate. The tool could also identify top residents to consider for awards or honors. Finally, being able to identify top or bottom cohorts could lead to further research questions, such as "What technical, behavioral, or communication skills do top (or bottom) performers have in common?" Answers may help further define "gold standards" of performance.

Our study has several potential limitations. First, the number and type of evaluators needed to accurately estimate residents' overall skill is unclear but may depend on both the rating and ranking task and the distribution of ability in the resident population. If our primary intention is to accurately classify individuals of lower and higher ability, the clustering of the data suggests that we can do with fewer evaluator types; nonphysician evaluators could be excluded. However, studies aimed at quantifying the sources of measurement error in the resident evaluations (eg, ranking bias) are needed. Second, relatively little data support the validity of the summary ranking measures. To establish validity would require comparing aggregate global rankings with other criterion measures (eg, performance on standardized, simulated scenarios). A criterion validation study would represent the next step for research and further address the usefulness of our GRI. Third, an evaluation system that incorporated multiple constructs and associated items might have produced a more mixed picture. For example, residents perceived as performing well in 1 competency might be perceived as not performing well in others. We chose our global assessment because no well-established method exists to aggregate data across all possible competencies. We also chose it so that we could consistently identify low and high performers, whose specific skills can be further assessed. The number and type of specific skills considered useful and important by an academic program may help to generate additional global questions for ranking purposes.

Conclusion

We evaluated whether physicians' and nonphysicians' global rankings, gathered via an Internet-based application, could identify high and low performers for 2 sets of anesthesiology resident classes. The Internet is useful for delivering assessment tools to diverse groups of evaluators. Whereas summary rankings can discriminate between low and high performers, a detailed review of the skills of high performers can provide benchmarks to guide standard setting for other performance-measurement modalities, for example, management of simulated adverse events.[23] Scoring of both technical and nontechnical skills in simulation exercises could be then contrasted to the quintile ranking that a resident actually received. From a patient-safety perspective, however, identifying low-performing residents is equally vital. Although the global ranking instrument does not elucidate individual competencies of low performers, it identifies those whose specific skills need detailed evaluation.

References

1 Gray JD. Global rating scales in residency education. *Acad Med*. 1996;71:S55–S63.
2 Accreditation Council for Graduate Medical Education (ACGME) outcome project. Toolbox of assessment methods. A product of the joint initiative of the ACGME outcome project of the ACGME and ABMS. Version 1.1, 2000. http://www.acgme.org/Outcome/. Accessed on June 21, 2010.
3 Keynan A, Friedman M, Benbassat J. Reliability of global rating scales in the assessment of clinical competence of medical students. *Med Educ*. 1987;21:477–481.
4 Daelmans H, van der Hem-Stokroos H, Hoogenboom R, Scherpbier A, Stehouwer C, van der Vleuten C. Global clinical performance rating, reliability and validity in an undergraduate clerkship. *Neth J Med*. 2005;63:279–284.
5 Scheffer S, Muehlinghaus I, Froehmel A, Ortwein H. Assessing students' communication skills: validation of a global rating. *Adv Health Sci Educ*. 2008;13:583–592.
6 Newble D. Techniques for measuring clinical competence: objective structured clinical examinations. *Med Educ*. 2004;38:199–203.
7 Tetzlaff J. Assessment of competency in anesthesiology. *Anesthesiology*. 2007;106:812–825.
8 Klass D. Assessing doctors at work—progress and challenges. *N Engl J Med*. 2007;356:414–415.
9 Epstein RM. Assessment in medical education. *N Engl J Med*. 2007;356:387–396.
10 Gaba DM, Howard SK, Flanagan B, Smith BE, Fish KJ, Botney R. Assessment of clinical performance during simulated crises using both technical and behavioral ratings. *Anesthesiology*. 1998;89:8–18.
11 Melnick DE, Asch DA, Blackmore DE, Klass DJ, Norcini JJ. Conceptual challenges in tailoring physician performance assessment to individual practice. *Med Educ*. 2002;36:931–935.
12 Solomon D, Szauter K, Rosebraugh C, Callaway M. Global ratings of student performance in a standardized patient examination: is the whole more than the sum of the parts? *Adv Health Sci Educ*. 2000;5:131–140.
13 Domingues RCL, Amaral E, Zeferino AMB. Global overall rating for assessing clinical competence: what does it really show? *Med Educ*. 2009;43:883–886.
14 Tabuenca A, Welling R, Sachdeva AK, et al. Multi-institutional validation of a web-based core competency assessment system. *J Surg Educ*. 2007;64:390–394.
15 Scott RS, Lind DS. An internet-based tool for evaluating third-year medical student performance. *Am J Surg*. 2003;185:211–215.
16 Bennett AJ, Arnold LM. Use of a computerized evaluation system in a psychiatry clerkship. *Acad Psychiatry*. 2004;28:197–203.
17 Wenrich MD, Carline JD, Giles LM, Ramsey PG. Ratings of the performances of practicing internists by hospital-based registered nurses. *Acad Med*. 1993;68:680–687.
18 Davis JDM. Comparison of faculty, peer, self, and nurse assessment of obstetrics and gynecology residents. *Obstet Gynecol*. 2002;99:647–651.
19 Johnson DM, Cujec BM. Comparison of self, nurse, and physician assessment of residents rotating through an intensive care unit. *Crit Care Med*. 1998;26:1811–1816.
20 Norcini JJ. Peer assessment of competence. *Med Educ*. 2003;37:539–543.
21 Violato C, Lockyer JM, Fidler H. Assessment of pediatricians by a regulatory authority. *Pediatrics*. 2006;117:796–802.
22 Tlayjeh H, Arabi YM, Al-Dorzi HM, et al. Assessment of critical care physicians using a subjective global rating scale of the ACGME competencies versus a simulation-based crisis resource management scale. *Am J Respir Crit Care Med*. 2010;181:A1663–A1667.
23 Gaba DM, Howard SK, Flanagan B, Smith BE, Fish KJ, Botney R, Assessment of clinical performance during simulated crises using both technical and behavioral ratings. *Anesthesiology*. 1998;89(1):8–18.