Determining Resident Clinical Performance

Getting Beyond the Noise

Keith Baker, M.D., Ph.D.*

ABSTRACT

Background: Valid and reliable (dependable) assessment of resident clinical skills is essential for learning, promotion, and remediation. Competency is defined as what a physician can do, whereas performance is what a physician does in everyday practice. There is an ongoing need for valid and reliable measures of resident clinical performance.

Methods: Anesthesia residents were evaluated confidentially on a weekly basis by faculty members who supervised them. The electronic evaluation form had five sections, including a rating section for absolute and relative-to-peers performance under each of the six Accreditation Council for Graduate Medical Education core competencies, clinical competency committee questions, rater confidence in having the resident perform cases of increasing difficulty, and comment sections. Residents and their faculty mentors were provided with the resident's formative comments on a biweekly basis.

Results: From July 2008 to June 2010, 140 faculty members returned 14,469 evaluations on 108 residents. Faculty scores were pervasively positively biased and affected by idiosyncratic score range usage. These effects were eliminated by normalizing each performance score to the unique scoring characteristics of each faculty member (Z-scores). Individual Z-scores had low amounts of performance information, but signal averaging allowed determination of reliable performance scores. Average Z-scores were stable over time, related to external measures of medical knowledge, identified residents referred to the clinical competency committee, and increased when performance improved because of an intervention.

Conclusions: This study demonstrates a reliable and valid clinical performance assessment system for residents at all levels of training.

What We Already Know about This Topic

- Evaluating clinical performance of resident trainees is essential to education, but the validity of evaluation methods has been questioned.

What This Article Tells Us That Is New

- In a 2-yr period, more than 14,000 electronic evaluations were submitted by faculty. Significant grade inflation could be removed by normalizing scores to each faculty member, yielding a more reliable and valid assessment of resident clinical skills.

RELIABLE measures of clinical performance are needed to enhance and direct learning, determine which trainees are ready for advanced training, and identify which are in need of remediation.[1,2] Unfortunately, evaluations of resident clinical performance suffer from a number of limitations,[3–5] such as trainees not being directly observed,[3] faculty leniency and grade range restriction,[6–8] concerns about validity of what is being assessed,[9–11] and the finding that even highly valid tests of medical knowledge may not[12,13] or may only modestly[14–17] predict competence in patient care. There are also issues of generalizability because Objective Structured Clinical Examinations (OSCEs)[18] and simulation-based examinations[19,20] sample only a subset of the domain of interest, and performance may not generalize to different circumstances.[10,21,22] Further-

◇ This article is featured in "This Month in Anesthesiology." Please see this issue of ANESTHESIOLOGY, page 9A.

◆ This article is accompanied by an Editorial View. Please see: Schwartz AJ: In the eyes of the beholder! ANESTHESIOLOGY 2011; 115:681–2.

* Assistant Professor of Anesthesia, Harvard Medical School; Assistant Anesthetist, Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital, Boston, Massachusetts.

Received from Massachusetts General Hospital, Boston, Massachusetts. Submitted for publication January 17, 2011. Accepted for publication June 8, 2011. Support was provided solely from institutional and/or departmental sources.

One or more authors of this peer-reviewed article have been supported by FAER. In conjunction with the FAER 25th anniversary, articles and editorials in the ANESTHESIOLOGY October 2011 issue celebrate the accomplishments of FAER. For additional information, visit www.FAER.org.

Address correspondence to Dr. Baker: Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital, 55 Fruit Street, Jackson 4, Boston, Massachusetts 02114. khbaker@partners.org. Anesthesiology 2011; 115:862–78

more, even when faculty members observe the same clinical performance, they may disagree about their observations[23] or what constitutes an acceptable performance[24] or response to a situation.[25] Lastly, and of considerable importance, is that physicians' scores on high-stakes OSCEs may not predict what they do in actual practice.[26] Thus, measures of competence (what a physician can do) may not relate to performance (what a physician actually does in everyday practice).[27,28]

This article describes an approach to assessing anesthesia resident clinical performance using the Accreditation Council for Graduate Medical Education (ACGME) core competency framework, is based on what residents do in everyday practice, depends on direct observation, uses many different evaluators representing a wide range of situations, is linked to written formative feedback, and yields a large number of evaluations. It was hypothesized that clinical performance scores could be corrected for faculty member leniency (positive bias) and idiosyncratic grade range usage and then averaged to yield a normalized resident performance metric that was valid and that distinguished clinical performance levels with known degrees of statistical confidence. The clinical performance metric is stable over time, reliably identifies low performers, detects improvement in performance when an educational intervention is successful, is related to an external measure of medical knowledge, and identifies poor performance due to a wide variety of causes.

Materials and Methods

The Massachusetts General Hospital Institutional Review Board waived the need for informed consent and classified this study as exempt.

Evaluation Instrument and Evaluation Process

The department's Education Committee created an initial evaluation instrument that was sent to the full faculty for comment. Faculty input was incorporated, and an updated version was sent to all residents for additional comment. Resident feedback was incorporated, and the Education Committee created a final version of the instrument. The resident evaluation form has five distinct sections (appendix) and is confidential for the evaluator.

Absolute/Anchored ACGME Core Competencies Section. The six ACGME core competencies are used, but patient care is divided into cognitive and technical sections yielding seven competency scores. The absolute/anchored scale uses a Likert scale (1–7) with descriptors of how much help the resident needed relating to each competency. A score of 5 was defined as performing independently and without the need for help.

Relative ACGME Core Competencies Section. The relative scale asks how the resident performed compared with other residents in residency in the same training year. The relative scale uses a Likert scale (1–5) with descriptors of how the resident performed compared with peers. A score of 3 is defined as performing at peer level (average) compared with other Massachusetts General Hospital anesthesia residents in the same clinical anesthesia year (CA-year).

Comments. Comment boxes occur frequently within the form (after each core competency, specific strengths, specific areas for improvement, and the clinical competency committee [CCC] section).

CCC Section. Five statements relating to essential competency attributes are listed. Each has a yes or no answer and any "yes" is considered concerning.

Faculty Member Confidence Section. Faculty members indicate their willingness to let the resident provide independent and unsupervised care for each of eight cases of increasing difficulty.

The faculty was formally educated on this instrument during conferences and faculty meetings, but not all faculty members attended the sessions.

Who Evaluates Whom. The electronic anesthesia system automatically determines which residents are supervised by which faculty members during the previous week (Sunday–Saturday). Duplicate interactions are collapsed into a single request for evaluation. Faculty members are permitted to submit additional evaluations at any time. For rotations that do not use the electronic anesthesia system (intensive care unit, preoperative clinic), matches are created by hand. The list of resident-faculty pairs is automatically sent to the New Innovations (New Innovations, Inc, Uniontown, OH) web site, which generates an electronic evaluation for each unique interaction. Faculty members are sent a link *via* electronic mail containing the evaluations and are automatically sent reminder e-mails if they do not complete the evaluations within a week. Overall compliance is tracked, with a target of 60% for each faculty member. Noncompliant faculty members are contacted and encouraged to complete outstanding evaluations. Completed evaluations are downloaded from the New Innovations web site as Excel spreadsheets (Version 2003, Microsoft, Redmond, WA). Raw data are imported into Access (Version 2003, Microsoft) for analysis.

Z-scores

Z-scores normalize a single resident evaluation to the unique scoring attributes of the faculty member providing the evaluation. Evaluations submitted within a specified date window are used to determine the characteristics of each faculty member's scoring attributes. Z-scores were determined using absolute/anchored core competency scores (Z_{abs}), relative-to-peers core competency scores (Z_{rel}), or case confidence scores (Z_{conf}). Each faculty member's Likert scores were used to determine his or her personal mean and SD for each CA-year. Individual resident Z-scores were calculated as:

$$Z = \frac{\text{(Resident Score [CA-year]} - \text{Faculty Member Mean [CA-year])}}{\text{(Faculty Member SD [CA-year])}}$$

Resident Score (CA-year) is the Likert Score assigned to a particular resident by a faculty member. When more than one core competency section is included, the average of the Likert scores from the selected core competencies is used.

Faculty Member Mean (CA-year) is the mean Likert score given to residents of a similar CA-year by this faculty member.

Faculty Member SD (CA-year) is the SD of Likert scores given by this faculty member to residents of this CA-year.

Z-scores provide a measure of distance from the grader's mean score in terms of SD units. For example, a Z-score of −0.5 means that the faculty member scored the resident one half SD less than he or she normally scores residents of this same CA-year. Z-scores are essentially effect sizes because they are differences normalized by the SD. Any combination of core competencies can be used in the calculation of a Z-score. When core competencies are not mentioned, a Z-score refers to an average based on all of the core competencies. Faculty member confidence data were converted to Z-scores by first determining the breakpoint at which the faculty member converted from "yes" to "no" along the sequence of eight graded cases. For example, if a faculty member said yes to the first three cases and no for the remaining five cases, the breakpoint would be 3. This allows the determination of the mean and SD of the breakpoints for each faculty member for each CA-year.

In-training Examination Z-scores

Z-scores for the American Society of Anesthesiologists/American Board of Anesthesiology In-Training Examination (ITE) (Z_{ITE}) were computed for each resident by first subtracting the resident's individual ITE scores from his or her Massachusetts General Hospital residency class mean (CA-year–matched classmates) and then dividing by the class SD.

Statistical Analysis

Statistical results were determined using StatsDirect Version 2.6.6 (StatsDirect Ltd., Cheshire, United Kingdom), Excel (Version 2003), SAS Version 9.2 (SAS Institute, Cary, NC), or Origin Version 7.5 SR4 (OriginLab, Northampton, MA). Effect sizes were determined by Cohen *d* and provide a measure of the size of a difference compared with the variation in the data. Effect sizes are classified as small (Cohen $d = 0.2$), medium (Cohen $d = 0.5$), or large (Cohen $d = 0.8$).[29] Regression analyses are characterized by *r* and r^2 (explained variance) values along with the number of data points used in the regression. Slopes were determined using linear regression. Slopes were compared using a Z-test statistic.[30] Repeat tests on the same sample are compared with paired *t* tests. Independent samples are compared with unpaired *t* tests assuming unequal group variance. Single-sample *t* tests compared a specified reference value to a sample of values. Sample variances were compared using an F test. Chi-square analysis was used for categorical data and Yates' correction was applied if expected frequencies were less than 10. Scores for relative ACGME core competencies were compared in a linear mixed model (LMM) with fixed effects for resident year (CA-1, -2, or -3); length of training within year at the time of the evaluation to accommodate improvement in scores over the course of training; and the interaction between resident year and length of training, random participant- and faculty-specific intercepts, and variance heterogeneity by faculty member. Nonlinearity in the trends over length of training was assessed using a cubic spline, but the fit was not improved based on Akaike information criterion. Point and interval estimates from this analysis were compared with results obtained from analyses of Z-scores. LMM estimates of participant-specific CIs were roughly 20% wider and more variable than matched Z-score estimates, but inference for comparisons among resident years was unchanged. *P* values were two-sided. The term "bias" is used throughout the study to denote the systematic tendency to assign performance scores that are higher than is normatively possible. With this particular usage, bias implies leniency. The terms "reliable" or "reliably" refer to dependable findings. With this usage, a score with a narrow 95% CI would be called reliable.

Results

Completed Evaluations

Between July 1, 2008, and June 30, 2010, 14,469 evaluations were submitted. This represents an overall (all requested, all returned) compliance rate of 49%. Evaluations were submitted by 140 different faculty members, who entered at least 5 evaluations on a total of 108 different residents, who each had at least 10 evaluations. There were 5,404 CA-1, 4,319 CA-2, and 4,746 CA-3 resident evaluations. On average, each CA-1, CA-2, and CA-3 resident received 101, 70, and 73 evaluations, respectively. On average, each CA-1, CA-2, and CA-3 resident was evaluated by 49, 40, and 41, respectively, different faculty members. Comments were entered on 59.1% of all returned evaluations. Comments averaged 225 ± 209 characters.

Faculty Members Characterize Resident Performance with a Positive Bias

The relative performance Likert scale defined 3 as "peer average" for each CA-year. This is explicitly stated on each evaluation form. The average relative score assigned for all core competencies by each faculty member contributing at least 10 evaluations was determined using all data. The average faculty member assigned a relative score of 3.36, 3.51, and 3.68 to CA-1, CA-2, and CA-3 residents, respectively. Histograms of the average relative score assigned by each faculty member by CA-year are shown in fig. 1. Using the expected value of 3.00 and the known SD of the faculty score distributions yields effect sizes for the bias of 0.91, 1.15, and 1.41 ($P < 0.001$ by single-sample *t* test, all cohorts) for scores assigned to the CA-1, CA-2, and CA-3 residents, respectively. These are large effects because average scores are approximately 1 SD above the expected value of 3.00.

Faculty members also increase their bias as they score more senior residents. For faculty members who provided both CA-1 and CA-2 evaluations, average CA-2 relative scores were higher (CA-1 = 3.36 *vs.* CA-2 = 3.48, N = 78

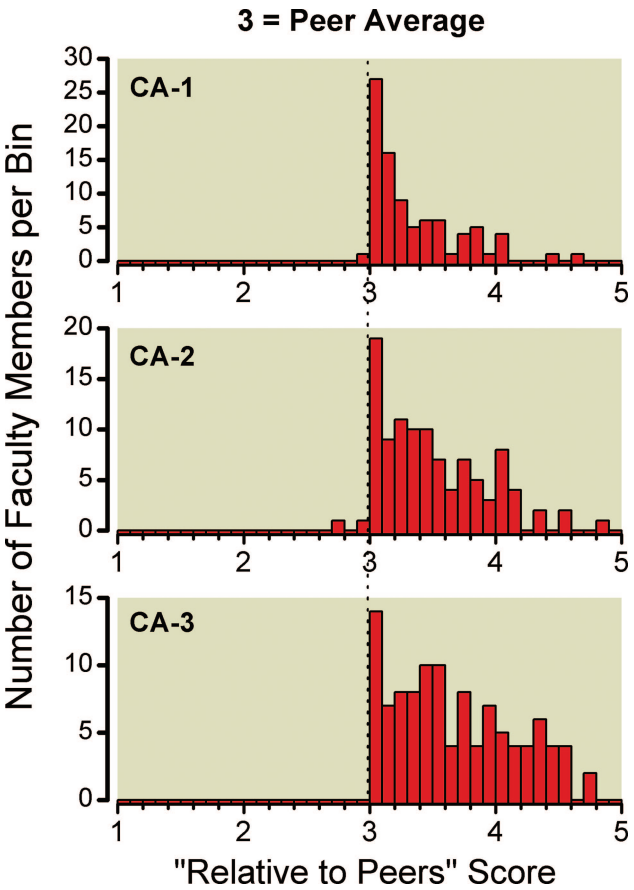

Fig. 1. Faculty members assign positively biased relative-to-peers scores. Histograms show counts, by clinical anesthesia (CA) year, of faculty members who assigned similar average relative-to-peers scores to residents. Average scores can range from 1 to 5, and 3 is defined as peer average. Bin widths are 0.1 score unit, and all faculty with an average score in that bin are counted. Counts were made for faculty members who submitted 10 or more evaluations per CA-year (CA-1: 88 faculty members, 4,630 evaluations; CA-2: 105 faculty members, 3,663 evaluations; CA-3: 110 faculty members, 4,034 evaluations).

faculty, $P < 0.001$ by paired t test). For faculty members who provided both CA-2 and CA-3 evaluations, average CA-3 relative scores were higher (CA-2 = 3.52 *vs.* CA-3 = 3.72, N = 97 faculty members, $P < 0.001$ by paired t test). For faculty members who provided both CA-1 and CA-3 evaluations, average CA-3 relative scores were higher (CA-1 = 3.37 *vs.* CA-3 = 3.70, N = 79 faculty members, $P < 0.001$ by paired t test).

Bias Varies by Faculty Member

All faculty members have their own amount of bias. Their average relative-to-peers scores are widely distributed (SD = 0.46, fig. 1). Scores from a relatively unbiased faculty member are compared with scores from a more biased faculty member in figure 2A. In addition to the variation in bias, faculty members also use different amounts of the score range. The used score range can be quantified by the SD of the scores given by each faculty member. Figure 2B shows a

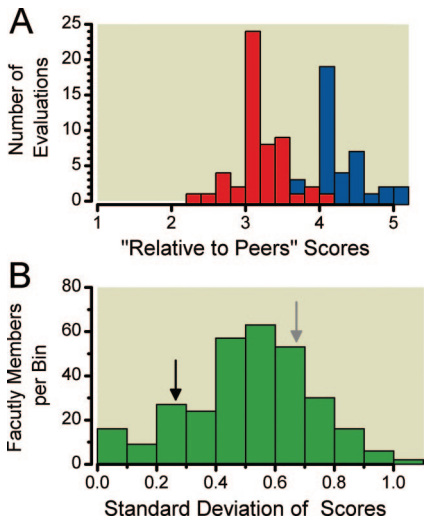

Fig. 2. Faculty members differ in how much they inflate scores. A histogram of all relative-to-peers scores from a relatively unbiased faculty member (*red bins,* N = 53 evaluations of residents in their third year of clinical anesthesia [CA-3]) is compared with a more biased faculty member (*blue bins,* N = 42 CA-3 evaluations) (*A*). Faculty members differ in how they use the available score range. The SD in score assignment was determined for each faculty member having 10 or more evaluations per CA-year. Bins are 0.1 SD units wide, and all faculty members with an average SD in that bin were counted. Data are from 88 faculty members with CA-1 data, 105 faculty members with CA-2 data, and 110 faculty members with CA-3 data. The *black arrow* denotes a faculty member with a SD of 0.26 for CA-1 scores, and the *gray arrow* denotes a faculty member with a SD of 0.68 for CA-1 scores (*B*).

histogram of SD for all faculty members having 10 or more evaluations for each of the CA-years (SD = 0.22). Faculty members use different amounts of the score range, as demonstrated by the lower average SD in scores given by one faculty member (SD = 0.26, N = 117 evaluations, CA-1 year, *black arrow* fig. 2B) compared with the higher average SD in scores given by another faculty member (SD = 0.68, N = 104 evaluations, CA-1 year, *gray arrow* fig. 2B).

Z_{rel} Scores Correct for Individual Faculty Member Bias and Unique Score Range Use

Because faculty members are biased to various degrees (fig. 1) and they each use different amounts of the score range (fig. 2B), a Z-score transformation was applied to the relative-to-peers scores (see Methods). Each faculty member's Z_{rel} scores thus have an overall mean of 0.0 and SD of 1 for each CA-year. All Z_{rel} scores for all residents were averaged (N = 13,639 evaluations), and the grand mean was 0.00000 with SD of 0.98623.

When a Faculty Member Evaluates the Same Resident on Two Occasions, the First Z_{rel} Score Predicts Only a Small Amount of the Variance in the Second Z_{rel} Score

Z_{rel} scores were determined for the first and second occasions when a faculty member evaluated the same resident more

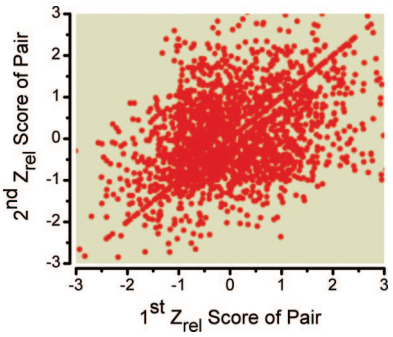

Fig. 3. Each Z_{rel} score has only a modest amount of clinical performance information. The first Z_{rel} score assigned to a resident by a faculty member is plotted against the second Z_{rel} score assigned to the same resident by the same faculty member. Each of the 3,509 points is a unique resident–faculty member pairing. The first Z_{rel} score predicts 23.1% of the variance in the second Z_{rel} score; 1.6% of the Z_{rel} scores lie outside the plot limits and are not shown.

than once. This resulted in 3,509 unique Z_{rel} score pairings. A regression analysis demonstrated that the first Z_{rel} score explained 23.1% of the variance of the second Z_{rel} score (N = 3,509 pairs, $r = 0.48$, $r^2 = 0.231$, $P < 0.001$). A plot of unique Z_{rel} score pairs demonstrates significant scatter in the data (fig. 3).

Signal Averaging Reveals Reliable Performance Scores

Because there is significant "noise" in each Z_{rel} score, any single Z_{rel} score will not provide a dependable assessment of resident clinical performance. However, averaging noisy signals will cause accumulation of the real signal while averaging out the noise component. Figure 4A demonstrates how sequential Z_{rel} scores yield a running average with a tighter and tighter nominal 95% CI as more signals (Z_{rel} scores) are averaged. A histogram of Z-scores for this individual shows how Z-scores are distributed about the mean (fig. 4B).

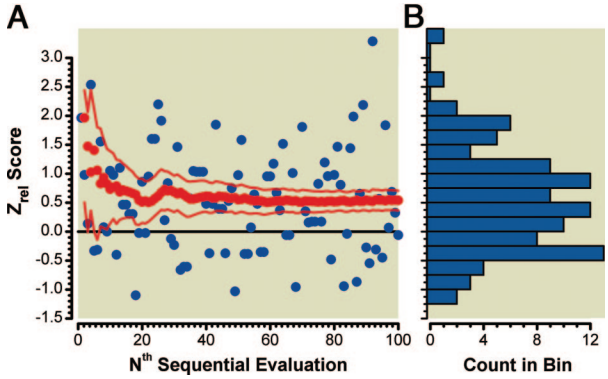

Fig. 4. Signal averaging reveals a reliable clinical performance score from a noisy background. The first 100 sequential Z_{rel} scores are shown for a single resident (*blue circles*). The running average and the upper and lower 95% CIs on the running average are shown by the *red filled circles* and *red lines*, respectively (*A*). Z-scores are distributed broadly about the mean. Z-scores (N = 100) from the same resident are displayed as a histogram with a bin width of 0.25 (*B*).

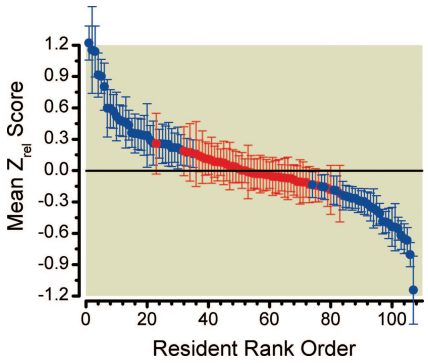

Fig. 5. Residents differ in their relative clinical performance. All data were used to determine mean Z_{rel} scores for each resident having 20 or more evaluations. Error bars are the 95% CI on the mean. Residents with a mean that is reliably above or below 0 are shown in *blue*. Residents with a mean that is not reliably different from 0 are shown in *red*.

Z_{rel} Scores Reliably Differentiate Relative Performance

All Z_{rel} scores were used to determine each resident's mean Z_{rel} and 95% CI. Of the 107 residents with 20 or more Z_{rel} scores, 32 (30%) were reliably above average, 46 (43%) were not reliably different from average, and 29 (27%) were reliably below average (fig. 5). When overall resident performance was determined using absolute data (Z_{abs}) or case confidence data (Z_{conf}), the different metrics yielded performance measures that were similar to Z_{rel}. A resident's mean Z_{abs} was related to his or her mean Z_{rel} ($r = 0.91$, $r^2 = 0.83$, N = 105 residents, $P < 0.001$). A resident's mean Z_{conf} was related to his or her mean Z_{rel} ($r = 0.57$, $r^2 = 0.33$, N = 105 residents, $P < 0.001$). The number of evaluations with usable Z_{conf} data were only 32.2% of the number with usable Z_{rel} data (13,639). The lower correlation of Z_{conf} with Z_{rel} was not attributable to a sampling bias because the correlation was unchanged when the correlation was determined using only forms containing both Z_{conf} and Z_{rel} data ($r = 0.58$, $r^2 = 0.34$, N = 105 residents, $P < 0.001$).

Average Z_{rel} Scores Determine Resident Performance as Well as a Sophisticated LMM

Average Z_{rel} scores and associated CIs do not take into account the repeated measures inherent in scoring the same resident on two or more occasions or scoring multiple residents by the same rater. To determine whether repeated measures were altering the estimates of resident clinical performance, average Z_{rel} scores (based on 20 or more samples) were compared with performance estimates determined using relative-to-peers data in a LMM. Z_{rel} scores provided a performance metric that was nearly identical to one determined using a LMM ($r = 0.96$, $r^2 = 0.92$, N = 107 residents, $P < 0.001$). The ratio of the resident variance component to residual variation was 27%. Thus, the repeated scores for a given resident are not fully independent., and the CIs determined by simple averaging of Z_{rel} scores will be narrower when repeated measures are included. The magnitude of this effect was determined by comparing CIs deter-

mined using Z_{rel} scores to those determined by the LMM. On average, the 95% CIs were 17.7% wider when determined using the LMM than when determined using Z_{rel} scores (N = 107 resident's 95% CIs, $P < 0.001$ by paired t test). The variance in the 95% CI was also higher when determined using the LMM (variance in Z_{rel} score 95% CI = 0.0016, variance in LMM 95% CI = 0.0031, $P < 0.001$ by F test).

There Is More Certainty in Determining Below-average Performances

When the SD of Z_{rel} is small, it indicates lower variation in the underlying Z_{rel} scores used to determine the mean. This leads to more certainty in the average score. When the SD of each resident's mean Z_{rel} score was regressed against the mean Z_{rel} score for the 107 residents with 20 or more Z_{rel} scores, the regression showed that the lower the Z_{rel}, the lower the SD ($r = 0.60$, $r^2 = 0.37$, N = 107 residents, $P < 0.001$). Thus, there is less variation in individual Z_{rel} scores for the lowest-performing residents than for the highest-performing residents. The number of evaluations submitted each month per resident did not differ between residents whose mean Z_{rel} was above 0 (9.88 evaluations per month, N = 543 resident-months) and those whose mean Z_{rel} was below 0 (9.98 evaluations per month, N = 696 resident-months) (unpaired t test, $P = 0.67$).

Z_{rel} Scores Are Stable When No Performance Interventions Occur

The temporal stability of each resident's Z_{rel} score was assessed by comparing his or her average Z_{rel} score during one 6-month period with the average Z_{rel} score 1 yr later during another 6-month period. All resident's having 15 or more evaluations during both 6-month periods (Period 1: October 1, 2008–March 31, 2009, Period 2: October 1, 2009–March 31, 2010) and who did not receive a performance intervention from the CCC were included. Forty-seven residents met these inclusion criteria. There was a strong relationship between the Z_{rel} scores from Period 1 and subsequent Z_{rel} scores from Period 2 ($r = 0.75$, $r^2 = 0.56$, N = 47 residents, $P < 0.001$, fig. 6). When the single outlier resident was removed, the relationship was strengthened ($r = 0.81$, $r^2 = 0.71$, N = 46 residents, $P < 0.001$).

Z_{rel} Scores for Medical Knowledge Are Related to an Independent Metric of Medical Knowledge: The American Society of Anesthesiologists/American Board of Anesthesiology ITE

Z_{rel} scores based solely on the core competency of Medical Knowledge ($Z_{rel,MK}$) were compared with the American Society of Anesthesiologists/American Board of Anesthesiology ITE examination. There were three cohorts of residents having both $Z_{rel,MK}$ scores and same-year ITE Z-scores (Z_{ITE}) (see Methods). The 2008 ITE was held in July. The 2009 and 2010 ITEs were held in March. The average $Z_{rel,MK}$ score for each resident was determined using evaluations submitted in the months after the exam (March through June). For each cohort, faculty member reference data were determined using their scores from the corresponding academic year (July–June). The 2008, 2009, and 2010 $Z_{rel,MK}$ scores were significantly related to the independently determined Z_{ITE} scores for each year examined (2008: $r = 0.38$, $r^2 = 0.14$, N = 71 residents, $P = 0.001$; 2009: $r = 0.33$, $r^2 = 0.12$, N = 76 residents, $P = 0.002$; 2010: $r = 0.30$, $r^2 = 0.09$, N = 69 residents, $P = 0.01$).

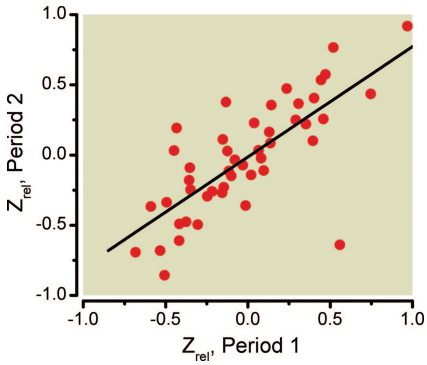

Fig. 6. Z_{rel} scores are stable over time when no interventions occur. Z_{rel} scores are shown for the 47 residents having no performance interventions and who had 15 or more evaluations in both Period 1 (October 1, 2008–March 31, 2009) and 1 yr later in Period 2 (October 1, 2009–March 31, 2010). The fitted line includes all data points ($r = 0.75$, $r^2 = 0.56$, N = 48, $P < 0.001$).

Z_{rel} Scores Independently Predict Referral to the CCC

Before the implementation of the new evaluation system, a number of residents had been referred to the CCC. The process leading to referral was multifactorial and included verbal communication, concerning written rotation evaluations, and electronic mail messages describing concerning performance. Once the Z_{rel} score system was functional, the system was used to see of it would identify residents who had been independently referred to the CCC. Residents with a Z_{rel} score greater than 0 were infrequently referred to the CCC (1 referred and 36 not). Residents with a Z_{rel} score of 0 or less were more often referred to the CCC (19 referred and 25 not). A Z_{rel} score of 0 or less was associated with an odds ratio of 27 in favor of being referred to the CCC ($P < 0.001$, two-tailed, chi-square with Yates' correction).

Z_{rel} Scores Predict CCC Flag Density of Below-average Performers

The evaluation form has five questions from the CCC that raise concern if answered "yes." CCC flag density is the fraction of evaluations having any of the CCC questions answered yes. For residents whose mean Z_{rel} score was less than 0, there was a strong inverse relationship between Z_{rel} score and CCC flag density ($r = 0.90$, $r^2 = 0.82$, N = 57 residents, $P < 0.001$). For residents whose mean Z_{rel} score was 0

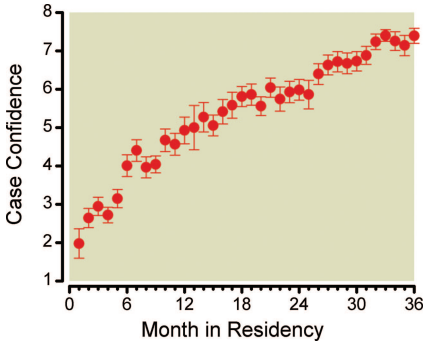

Fig. 7. Confidence increases as residency progresses. The mean maximum confidence (defined as the most advanced case the faculty has confidence in having the resident perform in an unsupervised fashion) is shown for all residents for all 36 months of residency (N = 5,006 evaluations). Confidence rises throughout residency but rises fastest during the first 12 months. The Y axis spans the case complexity used in the evaluation form: case 1 is relatively easy and case 8 is extremely challenging. Error bars are the 95% CI on the mean.

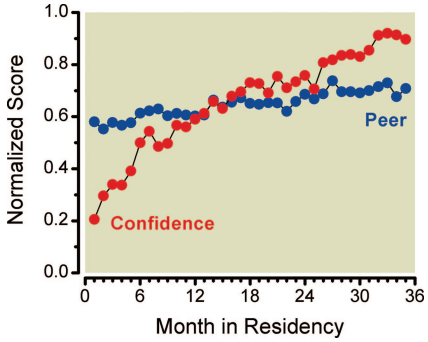

Fig. 8. Confidence scores increase faster than relative-to-peers scores as residency progresses. The average normalized confidence scores (*red*) and average normalized relative-to-peers scores (*blue*) are shown for each month of residency. Likert scores were normalized to a 0–1 scale, where 0 is the minimum and 1 is the maximum attainable score. Only evaluations having both a usable confidence score and a relative-to-peers score were included (N = 4,892). The overall slopes of the two data sets are different ($P < 0.001$, Z-test statistic).

or greater, there was no relationship between Z_{rel} score and CCC flag density ($r = 0.24$, $r^2 = 0.06$, N = 51 residents, $P = 0.10$).

Faculty Confidence in Having Residents Provide Unsupervised Care Increases as Residency Progresses

Faculty members provide a measure of their confidence in having the resident independently perform a series of eight cases of increasing difficulty. Of the evaluations completed, 5,006 had scores allowing a meaningful measure of when confidence was lost (see Methods). Confidence increased as months in residency increased (fig. 7). Confidence increased most rapidly during the first year of residency (slope = 0.25 cases/month, $r = 0.39$, $r^2 = 0.15$, N = 1,941 evaluations, $P < 0.001$) and slowed during the second year (slope = 0.09 cases/month, $r = 0.16$, $r^2 = 0.03$, N = 1,421 evaluations, $P < 0.001$) and third year (slope = 0.12 cases/month, $r = 0.27$, $r^2 = 0.07$, N = 1,644 evaluations, $P < 0.001$) of residency. The rate of increase in confidence was significantly higher during the first year of residency compared with either the second ($P < 0.001$, Z-test statistic) or third year ($P < 0.001$, Z-test statistic) of residency. The rate of increase was not different between the second and third years of residency ($P = 0.088$, Z-test statistic).

Confidence Scores Increase More than Relative Scores as Residents Become More Senior

Faculty members score residents increasingly above average as residents become more senior, although this is normatively impossible. If confidence scores rise disproportionately more than relative scores, this implies a real increase in actual performance and not just an increase in bias. Scores from evaluations containing both confidence and relative-to-peers data were normalized by their respective scale ranges such that 0.0 and 1.0 were the lowest and highest scores attainable. As residents progressed through residency, their normalized relative-to-peers scores increased (slope = 0.0044 normalized units/month, N = 4,982 evaluations, $P < 0.001$), as did their normalized confidence scores (slope = 0.018 normalized units/month, N = 4,982 evaluations, $P < 0.001$). The overall rate of increase was 4.0 times faster for the confidence data than for the relative-to-peers data ($P < 0.001$, Z-test statistic). Figure 8 shows the differential growth in normalized confidence scores compared with normalized relative-to-peers scores as residency proceeds.

A Performance Intervention Can Significantly Improve Z_{rel} Scores

Before this new system was used, a resident was referred to the program director using customary mechanisms. This resulted in an intervention in which performance issues were defined, written expectations were set forth, and consequences were defined. The program director, chair of the department, chair of the CCC, resident, and resident's mentor knew of the intervention. The faculty was otherwise unaware of the intervention. When the Z_{rel} score system became functional, previously collected data revealed that the faculty had independently assigned below-average Z_{rel} scores to this resident in the time leading up to the intervention ($Z_{rel} = -0.47$, upper bound on 95% CI did not include 0). The resident's Z_{rel} score increased significantly after the intervention ($Z_{rel} = 0.12$, 95% CI included 0, $P = 0.003$, unpaired t test). Figure 9 shows the Z_{rel} scores by month before and after the intervention. A second situation occurred after the Z_{rel} score system was in use. The CCC detected a resident with very low Z_{rel} scores, and a confidential educational intervention occurred. This included a written statement of specific concerns and expectations for improvement. The resident's Z_{rel} score for the 6 months leading up to

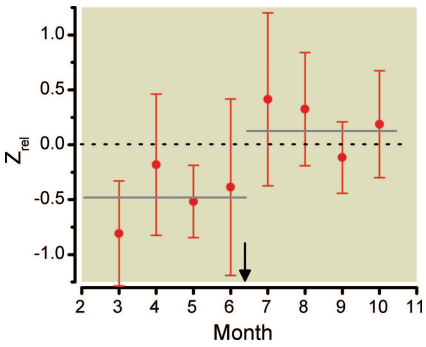

Fig. 9. Z_{rel} scores can increase after an education intervention. The mean monthly Z_{rel} scores for a single resident are shown for 4 months before and 4 months after an education intervention. The intervention occurred at the *arrow*. The resident's mean Z_{rel} scores for the 4 months before and after the intervention differ ($P = 0.003$) and are shown by the *gray lines*. Error bars are the 95% CI on the mean.

the intervention was well below average ($Z_{rel} = -0.66$, upper bound on 95% CI did not include 0). The average Z_{rel} score increased significantly for the 5 months after the educational intervention ($P < 0.001$, unpaired *t* test) and was no longer below average ($Z_{rel} = -0.02$, 95% CI included 0). Details and time courses of these two interventions are purposely left out to maintain anonymity of the residents.

Discussion

The Overall System

The reported resident evaluation system follows many of the recommendations found in the review of Williams *et al.*[3] and is consistent with the view that faculty members can, in aggregate, reliably and validly assess resident clinical performance.[31] The system is based on direct observation of clinical performance, has broad systematic sampling, uses multiple raters, uses a ACGME Core Competency construct, currently separates formative feedback and evaluative numbers, encourages weekly evaluation, occurs in a naturalistic setting with relatively unobtrusive observation, corrects for grade inflation (bias) and differential grade range use, is related to important metrics of performance such as high-stakes medical knowledge tests (ITE) and referral to a CCC, uses only five or seven rating choices per item, and specifies the meaning of ratings (table 1). A key finding in this study is that each Z_{rel} score has only limited clinical performance information. These noisy data are handled effectively by signal averaging many scores to create an overall clinical performance metric for each resident. The analysis includes CIs, which are helpful when using data for decision-making. CIs help distinguish meaningful differences in performance from differences that are uncertain. Uncertainty can be caused by too few evaluations or large variations in the scores themselves. The author used his department's previous competency system to identify residents in need of remediation while gaining comfort with the Z_{rel} score system. The Z_{rel} score system essentially has supplanted the department's previous system because it reliably detects all residents who have significant performance issues. Despite Z_{rel} scores being normalized values that do not contain absolute clinical competency information, the experience at the institution has shown repeatedly that a mean Z_{rel} score of approximately -0.5 (or less) signals the need for intervention (fig. 5). Residents with a mean Z_{rel} score of less than approximately -0.6 present a challenge, and those with scores less than approximately -0.8 may face serious performance issues necessitating significant intervention. Residents whose Z_{rel} score is so low that their upper 95% CI does not reach -0.5 are most concerning. Unless otherwise noted, individual Z_{rel} scores are based on the average of the ACGME core competency subscores after a recent review found that raters typically are unable to assess independently the six core competencies.[32] This process appears to be one of the most robust and extensive evaluation systems found in the medical education literature.

Table 1. Features of the Clinical Performance Evaluation System

Direct observation of clinical performance
Broad systematic sampling
Multiple raters
ACGME core competency framework
Separation of formative feedback and evaluative numbers
Encourages weekly evaluation
Occurs in a naturalistic setting with relatively unobtrusive observation
Corrects for grade inflation (bias) and differential grade range use
Relates to high stakes medical knowledge tests (ITE) and referral to a CCC
Uses only five or seven rating choices per item
Specifies meaning of ratings

ACGME = Accreditation Council for Graduate Medical Education; CCC = clinical competency committee; ITE = In-Training Examination.

Z-scores Correct for Biases

The relative-to-peers component of the evaluation system asks faculty members to score a resident's performance relative to his or her peer group (same CA-year within the same residency) for each competency. Nearly every faculty member provided scores that were well above average (fig. 1). This bias was exaggerated when faculty members evaluated more senior residents. The finding that normative performance scores are inflated into the "above average" range is an example of the "Lake Wobegon" effect, which is not unique to physicians.[33] Because of the unique use patterns by each faculty member, it became apparent that a normalization process was needed to recenter the scores and adjust for differing score range use. Z-scores accomplish both of these requirements. In addition, because bias increased with CA-year, faculty scores were normalized for each CA-year. The

Z-score transformation reduces the amount of construct-irrelevant variance[11,34–36] in the data. Z-scores can be averaged and compared in units of SD. The Z-score transformed data behave as expected with a grand mean of 0 and a SD of nearly 1.

A Single Z_{rel} Score Has Only a Small Amount of Clinical Performance 'Truth' Associated with It

A key finding of this study was the low correlation between first and second Z_{rel} scores when a faculty member evaluated the same resident on two occasions (fig. 3). This indicates at most a modest halo effect[37] because faculty member scores differ significantly between subsequent evaluations of the same resident. Overall, approximately 23% of the second performance score can be explained by the first performance score. This small component likely contains the actual performance measure. This leaves 77% of the score as noise or unexplained variance. The low correlation between first and second Z_{rel} scores may be partly attributable to the differences in the situations leading to each Z_{rel} score. Clinical performance is highly affected by the circumstances of the event. This concept is known as "context specificity"[38,39] and explains why performance on one OSCE station predicts only a modest amount of the performance on the exact same OSCE station when using a different standardized patient.[21] Essentially, people fail to adequately consider the role of the situation in determining behavior and performance.[39,40]

Signal Averaging Is the Key to Determining Clinical Performance

Noisy signals such as Z_{rel} scores are well handled by signal averaging, which reduces the noise and reveals the signal. Figures 4A and B display significant variation in Z_{rel} scores but a running average that converges on a "true" Z_{rel} score with a small error signal. This allows an estimate of overall relative performance to emerge from the noise. Because of repeated measures, the Z_{rel} score CIs of below-average performers typically reach statistical significance with a smaller number of evaluations than if an LMM had been used. Thus, the Z-score system will detect low performers sooner and enable educators to get them the help they need.

Do Z-scores Really Provide a Measure of Clinical Performance?

There are four lines of evidence supporting Z_{rel} scores as a measure of actual clinical performance. First, Z_{rel} scores determined using just the scores for medical knowledge ($Z_{rel,MK}$) were related to an independent determination of medical knowledge. The strength of the relationship indicates that $Z_{rel,MK}$ scores explain approximately 10–15% of the variance in ITE scores. Second, the likelihood of being referred to the CCC was independently related to mean Z_{rel} scores. Residents with a Z_{rel} score of 0 or less were referred to the CCC with an odds ratio of 27. The author's CCC now uses Z_{rel} scores to detect low performers. Third, as residents progress through residency, the normalized confidence scores increased 4.0 times faster than the normalized relative scores (fig. 8). If scores were simply related to progressive bias or construct-irrelevant variance,[35,36] the ratio of normalized confidence to normalized relative scores would remain constant. Fourth, CCC flag density, an independent measure of concern with clinical performance, is strongly related to lower Z_{rel} scores.

The finding that residents with higher average Z_{rel} scores have more variance in their Z_{rel} scores is intriguing. One explanation may be that it is difficult to consistently deliver an above-average performance, and this may add variance to their scores. It is also possible that the faculty have more agreement on what constitutes poor performance than what constitutes excellent performance.[31]

Why Are $Z_{rel,MK}$ Scores Only Slightly Related to ITE Scores?

A modest but real relationship was found between the Z_{rel} score assessment of medical knowledge and the ITE-based assessment of medical knowledge. Faculty members are unaware of residents' ITE scores except for those few that they mentor, so the correlation is not caused by the faculty's knowledge of residents' ITE scores. Although United States Medical Licensing Examination scores predict future standardized test results, such as ITEs,[12,16] they are poorly[16] or not at all[12] related to clinical performance. Even when the medical knowledge being tested is related to the actual clinical scenario of an OSCE, it hardly predicts performance on that OSCE.[21] Thus, weak correlations between $Z_{rel,MK}$ and ITE scores are expected and may be attributable to a number of factors. Faculty members may not actively probe residents to determine the true extent of their medical knowledge. Furthermore, when residents and faculty members interact, they are using practical or applied medical knowledge, as opposed to the theoretical medical knowledge tested by standardized examinations. Most medical decisions in natural settings have significant amounts of uncertainty, are prone to bias and cognitive errors,[41,42] and require significant amounts of judgment.[43] This is in sharp contrast to ITE questions, which have only one correct answer. There is a significant amount of research showing that cognitive ability (intelligence) is poorly or not related to the ability to avoid biased thinking.[44–46] Thus, the Z_{rel} assessment of medical knowledge may be an excellent proxy for day-to-day clinical decision-making and serve as a metric for what residents do in practice, an important measure.

Z-scores Are Stable Unless the Resident Is Coached onto a New Plane of Performance

The stability of Z_{rel} scores over the course of 1 yr is significant (fig. 6). The mean Z_{rel} score from the first time period explained 56% of the variance in the mean Z_{rel} score 1 yr later, indicating that scores generally are stable. Recent studies indicate that certain personality traits are related to better and

worse clinical performance.[47,48] If this is true, stability in relative clinical performance can be explained partially by the general stability of personality traits.[49]

Z-scores Change When a Resident's Performance Changes

If clinical performance is not malleable, there is little reason to provide feedback. This article provides two clear examples of clinical performance improvement associated with a feedback intervention. There are three important features found in these examples (see fig. 9 for one example). First, the Z_{rel} score system independently identified the resident. Second, the resident's Z_{rel} scores increased after the intervention without the faculty being aware of the intervention. This indicates that the faculty view performance for what it is and do not allow previous reputation to taint significantly the evaluation process. Third, it associates feedback and an educational intervention with improved clinical performance, a key role of residency.[50] It is likely that the evaluation system served to identify a performance problem and track its improvement. The educational interventions, in conjunction with developmental feedback, are what likely caused the performance improvement.

Is There a Particular Score Defining Adequate Performance?

When the residents have average Z_{rel} scores of less than approximately −0.3 and the 95% CI does not include 0 (*i.e.*, their performance is reliably below average), the author's CCC carefully examines the corresponding comments to determine the nature of the low performance. It has been found that there are many routes to low performance, including poor medical knowledge, low effort, unprofessional behavior, interpersonal and communication difficulties, poor motivation to improve, confidence in excess of competence, defensiveness, anxiety, low confidence, poor decision-making, and so forth. The comments are used to help develop educational interventions that target the area in need of improvement. Residents exhibiting noncognitive and nontechnical causes of low performance (such as low motivation for learning, defensiveness, anxiety, and so forth) are readily identified using this system. However, the underlying causes sometimes can be difficult to identity. The comments section usually provides strong hints to the cause but not always. In situations in which the precise noncognitive cause for low performance cannot be identified, outside learning specialists, psychiatrists, cognitive behavioral therapists, and personal coaches have been used. The results usually have been quite rewarding. Additional information is limited to protect the privacy of individual residents.

The ACGME has reframed residency training to focus on outcomes instead of process.[32] Despite this call, there are few outcomes that independently measure competency and fewer still that measure performance. Unfortunately, even when OSCEs or other highly reliable metrics are used to determine clinical competency, there is only a weak relationship with actual clinical performance.[21,26,51] This indicates the need for more naturalistic measures of performance,[2,5,28,52–55] such as the one described in this article. Once clinical performance becomes measurable, there remains the task of standard setting. Standard setting is largely context sensitive; for example, a physician deemed acceptable by today's standards may not be considered acceptable by future standards. Thus, normative standards still have an important role in determining adequacy of performance.[31,56]

Limitations of the Study

This study is limited by its inability to establish absolute performance levels. However, the relationship between relative and absolute performance appears to be real based on the ability of Z_{rel} scores to predict ITE scores and CCC referrals. Z_{rel} scores assume normally distributed data, and faculty member scores may not always be normally distributed. Individual Z_{rel} scores contain only a modest signal, so large sample sizes are required to attain reliable measures of clinical performance. The Z_{rel} system does not take into account repeated measures; however, using an LMM to correct for repeated measures did not significantly affect the estimates of clinical performance. Importantly, averaging Z_{rel} scores typically results in more narrow CIs than those determined using an LMM. This may result in earlier detection of poor performance. The LMM is an excellent tool but does not easily lend itself to practical use. The current study is receiving approximately one half of the evaluations requested. This means there is a risk of a sampling error. Many different faculty members contribute to each resident's Z_{rel} score, so it is unlikely that the error is large. Another limitation is the delay in requesting an evaluation. The delay is, on average, one half a week but can be as short as 1 day or as long as 1 week, depending on when during the previous week the interaction occurred. A more concerning delay occurs when faculty members delay completing the evaluation. This can amount to many weeks or even months. Currently, outstanding evaluations are deleted after 3 months.

This study demonstrates that when faculty members evaluate resident clinical performance in a naturalistic setting that encompasses a variety of clinical situations, they assign scores that suffer from significant grade inflation and varying degrees of grade-range usage. The unique grading characteristics of each faculty member were used to normalize the scores that each faculty member assigned. Resulting single Z_{rel} scores were shown to contain a modest amount of true clinical performance information. The low information content of single scores was largely circumvented by averaging many independent scores to arrive at a metric that was related to clinical performance measures, including referral to the CCC, medical knowledge scores (ITE scores), and growth in faculty confidence in allowing residents to undertake independent and unsupervised care of increasingly complex patients. The strength of the system is its ability to average out

irrelevant variance, which leaves a useful metric of clinical performance. The metric was stable over time. Although the metric is normalized and thus does not measure absolute clinical performance, it is able to detect poor clinical performance, which faculty members, in aggregate, appear to agree upon. When mean Z_{rel} scores are less than approximately −0.5, it signals the need to look into the cause(s) of the poor performance, and the comments section can help identify what can be done to improve performance. Two exemplar residents with low clinical performance scores each received an educational intervention based on the information contained in the comments sections, and both experienced significant improvement in performance after the intervention.

The author thanks the faculty members who spent time and effort evaluating residents and extends a special thanks to those who wrote comments aimed at improving resident performance. The author also thanks Eric A. Macklin, Ph.D. (Instructor, Harvard Medical School, Assistant in Biostatics, Department of Medicine, Massachusetts General Hospital, Boston, Massachusetts), for statistical advice.

References

1. Epstein RM, Hundert EM: Defining and assessing professional competence. JAMA 2002; 287:226–35
2. Epstein RM: Assessment in medical education. N Engl J Med 2007; 356:387–96
3. Williams RG, Klamen DA, McGaghie WC: Cognitive, social and environmental sources of bias in clinical performance ratings. Teach Learn Med 2003; 15:270–92
4. Holmboe ES: Faculty and the observation of trainees' clinical skills: Problems and opportunities. Acad Med 2004; 79: 16–22
5. Kogan JR, Holmboe ES, Hauer KE: Tools for direct observation and assessment of clinical skills of medical trainees: A systematic review. JAMA 2009; 302:1316–26
6. Albanese MA: Challenges in using rater judgements in medical education. J Eval Clin Pract 2000; 6:305–19
7. Margolis MJ, Clauser BE, Cuddy MM, Ciccone A, Mee J, Harik P, Hawkins RE: Use of the mini-clinical evaluation exercise to rate examinee performance on a multiple-station clinical skills examination: A validity study. Acad Med 2006; 81: S56–60
8. McManus IC, Thompson M, Mollon J: Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling. BMC Med Educ 2006; 6:42
9. Norman GR, Van der Vleuten CP, De Graaff E: Pitfalls in the pursuit of objectivity: Issues of validity, efficiency and acceptability. Med Educ 1991; 25:119–26
10. Hodges B, Regehr G, McNaughton N, Tiberius R, Hanson M: OSCE checklists do not capture increasing levels of expertise. Acad Med 1999; 74:1129–34
11. Downing SM: Threats to the validity of clinical teaching assessments: What about rater error? Med Educ 2005; 39: 353–5
12. Rifkin WD, Rifkin A: Correlation between housestaff performance on the United States Medical Licensing Examination and standardized patient encounters. Mt Sinai J Med 2005; 72:47–9
13. McGaghie WC, Cohen ER, Wayne DB: Are United States Medical Licensing Exam Step 1 and 2 scores valid measures for postgraduate medical residency selection decisions? Acad Med 2011; 86:48–52
14. Norcini JJ, Webster GD, Grosso LJ, Blank LL, Benson JA Jr: Ratings of residents' clinical competence and performance on certification examination. J Med Educ 1987; 62: 457–62
15. Tamblyn R, Abrahamowicz M, Dauphinee WD, Hanley JA, Norcini J, Girard N, Grand'Maison P, Brailovsky C: Association between licensure examination scores and practice in primary care. JAMA 2002; 288:3019–26
16. Hamdy H, Prasad K, Anderson MB, Scherpbier A, Williams R, Zwierstra R, Cuddihy H: BEME systematic review: Predictive values of measurements obtained in medical schools and future performance in medical practice. Med Teach 2006; 28:103–16
17. Holmboe ES, Wang Y, Meehan TP, Tate JP, Ho SY, Starkey KS, Lipner RS: Association between maintenance of certification examination scores and quality of care for medicare beneficiaries. Arch Intern Med 2008; 168:1396–403
18. Stillman PL, Swanson DB, Smee S, Stillman AE, Ebert TH, Emmel VS, Caslowitz J, Greene HL, Hamolsky M, Hatem C, Levenson DJ, Levin R, Levinson G, Ley B, Morgan GJ, Parrino T, Robinson S, Willms J: Assessing clinical skills of residents with standardized patients. Ann Intern Med 1986; 105: 762–71
19. Scavone BM, Sproviero MT, McCarthy RJ, Wong CA, Sullivan JT, Siddall VJ, Wade LD: Development of an objective scoring system for measurement of resident performance on the human patient simulator. ANESTHESIOLOGY 2006; 105:260–6
20. McIntosh CA: Lake Wobegon for anesthesia…where everyone is above average except those who aren't: Variability in the management of simulated intraoperative critical incidents. Anesth Analg 2009; 108:6–9
21. Norman GR, Tugwell P, Feightner JW, Muzzin LJ, Jacoby LL: Knowledge and clinical problem-solving. Med Educ 1985; 19:344–56
22. Savoldelli GL, Naik VN, Joo HS, Houston PL, Graham M, Yee B, Hamstra SJ: Evaluation of patient simulator performance as an adjunct to the oral examination for senior anesthesia residents. ANESTHESIOLOGY 2006; 104:475–81
23. Herbers JE Jr, Noel GL, Cooper GS, Harvey J, Pangaro LN, Weaver MJ: How accurate are faculty evaluations of clinical competence? J Gen Intern Med 1989; 4:202–8
24. Noel GL, Herbers JE Jr, Caplow MP, Cooper GS, Pangaro LN, Harvey J: How well do internal medicine faculty members evaluate the clinical skills of residents? Ann Intern Med 1992; 117:757–65
25. Ginsburg S, Regehr G, Lingard L: Basing the evaluation of professionalism on observable behaviors: A cautionary tale. Acad Med 2004; 79:S1–4
26. Rethans JJ, Sturmans F, Drop R, van der Vleuten C, Hobus P: Does competence of general practitioners predict their performance? Comparison between examination setting and actual practice. BMJ 1991; 303:1377–80
27. Rethans JJ, van Leeuwen Y, Drop R, van der Vleuten C, Sturmans F: Competence and performance: Two different concepts in the assessment of quality of medical care. Fam Pract 1990; 7:168–74
28. Rethans JJ, Norcini JJ, Barón-Maldonado M, Blackmore D, Jolly BC, LaDuca T, Lew S, Page GG, Southgate LH: The relationship between competence and performance: Implications for assessing practice performance. Med Educ 2002; 36:901–9
29. Cohen J: A power primer. Psychol Bull 1992; 112:155–9
30. Paternoster R, Brame R, Mazerolle P, Piquero A: Using the correct statistical test for the equality of regression coefficients. Criminology 1998; 36:859–66
31. Greaves JD, Grant J: Watching anaesthetists work: Using the professional judgement of consultants to assess the developing clinical competence of trainees. Br J Anaesth 2000; 84:525–33
32. Lurie SJ, Mooney CJ, Lyness JM: Measurement of the general competencies of the accreditation council for graduate med-

ical education: A systematic review. Acad Med 2009; 84: 301–9
33. Jawahar IM, Williams CR: Where all the children are above average: The performance appraisal purpose effect. Personnel Psychology 1997; 50:905–25
34. Messick S: Standards of validity and the validity of standards in performance assessment. Educational Measurement: Issues and Practice 1995; 14:5–8
35. Downing SM, Haladyna TM: Validity threats: Overcoming interference with proposed interpretations of assessment data. Med Educ 2004; 38:327–33
36. Haladyna TM, Downing SM: Construct-Irrelevant Variance in High-Stakes Testing. Educational Measurement: Issues and Practice 2004; 23:17–27
37. Bechger TM, Maris G, Hsiao YP: Detecting halo effects in performance-based examinations. Appl Psychol Meas 2010; 34:607–19
38. Eva KW, Neville AJ, Norman GR: Exploring the etiology of content specificity: Factors influencing analogic transfer and problem solving. Acad Med 1998; 73:S1–5
39. Eva KW: On the generality of specificity. Med Educ 2003; 37:587–8
40. Darley JM, Batson CD: 'From Jerusalem to Jericho': A study of situational and dispositional variables in helping behavior. J Pers Soc Psychol 1973; 27:100–8
41. Croskerry P: The importance of cognitive errors in diagnosis and strategies to minimize them. Acad Med 2003; 78:775–80
42. Norman GR, Eva KW: Diagnostic error and clinical reasoning. Med Educ 2010; 44:94–100
43. Holmboe ES, Lipner R, Greiner A: Assessing quality of care: Knowledge matters. JAMA 2008; 299:338–40
44. Stanovich KE, West RF: On the relative independence of thinking biases and cognitive ability. J Pers Soc Psychol 2008; 94:672–95
45. Stanovich KE, West RF: On the failure of cognitive ability to predict myside and one-sided thinking biases. Thinking & Reasoning 2008; 14:129–67
46. West RF, Toplak ME, Stanovich KE: Heuristics and biases as measures of critical thinking: Associations with cognitive ability and thinking dispositions. J Educ Psychol 2008; 100: 930–41
47. Reich DL, Uysal S, Bodian CA, Gabriele S, Hibbard M, Gordon W, Sliwinki M, Kayne RD: The relationship of cognitive, personality, and academic measures to anesthesiology resident clinical performance. Anesth Analg 1999; 88:1092–100
48. Merlo LJ, Matveevskii AS: Personality testing may improve resident selection in anesthesiology programs. Medical Teacher 2009; 31:e551–4
49. Loehlin JC, Martin NG: Age changes in personality traits and their heritabilities during the adult years: Evidence from Australian twin registry samples. Pers Individ Dif 2001; 30: 1147–60
50. Sachdeva AK: Use of effective feedback to facilitate adult learning. J Cancer Educ 1996; 11:106–18
51. Hoppe RB, Farquhar LJ, Henry R, Stoffelmayr B: Residents' attitudes towards and skills in counseling: Using undetected standardized patients. J Gen Intern Med 1990; 5:415–20
52. Ginsburg S, McIlroy J, Oulanova O, Eva K, Regehr G: Toward authentic clinical evaluation: Pitfalls in the pursuit of competency. Acad Med 2010; 85:780–6
53. Govaerts MJ, van der Vleuten CP, Schuwirth LW, Muijtjens AM: Broadening perspectives on clinical performance assessment: Rethinking the nature of in-training assessment. Adv Health Sci Educ Theory Pract 2007; 12:239–60
54. Berwick DM: Measuring physicians' quality and performance: Adrift on Lake Wobegon. JAMA 2009; 302:2485–6
55. Schuwirth L, van der Vleuten C: Merging views on assessment. Med Educ 2004; 38:1208–10
56. Norcini JJ: Setting standards on educational tests. Med Educ 2003; 37:464–9

Appendix

WEB-BASED RESIDENT EVALUATION
GENERAL FORM
RESIDENT EVALUATION BY STAFF

Evaluator: **Subject:**
Status:
Rotation:
Employer:

DATE OF THIS EVALUATION:

EVALUATION DESIGNATIONS

ABSOLUTE/ANCHORED COMPETENCY DESIGNATION

A competent physician (rating of 5, 6 or 7) performs independently in a fashion that is consistent with the standard of care in the United States today. Ratings of 5, 6 or 7 imply that a resident does not require attending supervision. Thus, ratings of 5, 6 or 7 imply that the resident is ready to leave the residency.

1 = needed **significant** attending assistance, input or correction
2 = needed **moderate** attending assistance, input or correction
3 = needed only **minimal** assistance, input or correction (emerging as competent)
4 = needed **very infrequent** assistance, input or correction (emerging as competent)
5 = performed in a **fully independent** manner, did not need any faculty, input or correction
6 = **able to serve as a consultant to other physicians,** able to defend all actions and decisions
7 = expert and able to serve as **a resource to fully trained anesthesiologists**
N/A = **Not able to evaluate** resident on this competency

RELATIVE PERFORMANCE DESIGNATION

This designation normalizes the resident's performance to other Massachusetts General Hospital residents who are at the same level of training.

1 = distinctly below peer level
2 = somewhat below peer level
3 = at peer level (most residents should be at this level)
4 = somewhat above peer level
5 = distinctly above peer level
N/A = **Not able to evaluate** resident on this competency

Sample items to consider for each Core Competency:

Medical Knowledge
Knows mechanism of actions of induction drugs, including primary side effects
Knows indications and complications of various monitoring devices
Knows physiology of pertinent organs systems
Knows medical diseases and implications for anesthetic plan

Patient Care
Designs and defends anesthetic plan
Shows appropriate vigilance, judgment and decision-making for perioperative events, including procedures
Develops contingency plans for foreseen and unforeseen outcomes

Practice-Based Learning Carries out post-operative checks with the intent to learn how to improve the care for subsequent patients
Critically examines decisions and actions for optimal performance
Uses evidence-based medicine to the extent available
Seeks out and adjusts performance according to feedback

Professionalism
Is fully prepared in the mornings
Acts in a manner consistent with a medical professional
Takes timely breaks
Demonstrates a good work ethic
Is aware of and attends to the goals and objective for the rotation
Carries out tasks that may not have direct personal gain (pre-ops for a colleague)

Interpersonal & Communication Skills
Interacts with patients and perioperative personnel in a caring and thoughtful fashion
Explains and defends decisions in a defensible and understandable form
Writes complete and insightful preoperative notes
Consults surgeons and attending anesthesiologist in a functional time frame

Systems-Based Practice
Recognizes and acts in a manner that acknowledges that they are part of a larger system and that patient care is based on this system. Caring for waitlist cases with a positive attitude is a manifestation of this understanding.
Recognizes ways to improve the system, even if it does not pertain to their case
Recognizes and follows HIPPA regulations
Recognizes and appropriately interfaces with infection control issues
Prepares cases for case conference and QA processes when appropriate

Medical Knowledge-Absolute/Anchored (1-7)

Needed Significant Assistance 1	Needed Moderate Assistance 2	Needed only Minimal Assistance 3	Needed Very Infrequent Assistance 4	Performed in a Fully Independent Manner 5	Able to Serve as a Consultant to Other Physicians 6	Expert and Able to Serve as a Resource to Fully Trained Anesthesiologists 7	N/A
○	○	○	○	○	○	○	○

Medical Knowledge-Relative (1-5)

Distinctly Below Peer Level 1	Somewhat Below Peer Level 2	At Peer level (Most residents should be at this level) 3	Somewhat Above Peer Level 4	Distinctly Above Peer Level 5	N/A
○	○	○	○	○	○

Comments

Remaining Characters: 5000

Patient Care - Cognitive - Absolute/Anchored (1-7)

Needed Significant Assistance 1	Needed Moderate Assistance 2	Needed only Minimal Assistance 3	Needed Very Infrequent Assistance 4	Performed in a Fully Independent Manner 5	Able to Serve as a Consultant to Other Physicians 6	Expert and Able to Serve as a Resource to Fully Trained Anesthesiologists 7	N/A
○	○	○	○	○	○	○	○

Patient Care - Cognitive - Relative (1-5)

Distinctly Below Peer Level 1	Somewhat Below Peer Level 2	At Peer level (Most residents should be at this level) 3	Somewhat Above Peer Level 4	Distinctly Above Peer Level 5	N/A
○	○	○	○	○	○

Comments

Remaining Characters: 5000

Patient Care - Technical - Absolute/Anchored (1-7)

Needed Significant Assistance 1	Needed Moderate Assistance 2	Needed only Minimal Assistance 3	Needed Very Infrequent Assistance 4	Performed in a Fully Independent Manner 5	Able to Serve as a Consultant to Other Physicians 6	Expert and Able to Serve as a Resource to Fully Trained Anesthesiologists 7	N/A
○	○	○	○	○	○	○	○

Patient Care - Technical - Relative (1-5)

Distinctly Below Peer Level 1	Somewhat Below Peer Level 2	At Peer level (Most residents should be at this level) 3	Somewhat Above Peer Level 4	Distinctly Above Peer Level 5	N/A
○	○	○	○	○	○

Comments

Remaining Characters: 5000

Practice-based learning - Absolute/Anchored (1-7)

Needed Significant Assistance 1	Needed Moderate Assistance 2	Needed only Minimal Assistance 3	Needed Very Infrequent Assistance 4	Performed in a Fully Independent Manner 5	Able to Serve as a Consultant to Other Physicians 6	Expert and Able to Serve as a Resource to Fully Trained Anesthesiologists 7	N/A
○	○	○	○	○	○	○	○

Practice-based learning - Relative (1-5)

Distinctly Below Peer Level 1	Somewhat Below Peer Level 2	At Peer level (Most residents should be at this level) 3	Somewhat Above Peer Level 4	Distinctly Above Peer Level 5	N/A
○	○	○	○	○	○

Comments

Remaining Characters: 5000

Professionalism - Absolute/Anchored (1-7)

Needed Significant Assistance 1	Needed Moderate Assistance 2	Needed only Minimal Assistance 3	Needed Very Infrequent Assistance 4	Performed in a Fully Independent Manner 5	Able to Serve as a Consultant to Other Physicians 6	Expert and Able to Serve as a Resource to Fully Trained Anesthesiologists 7	N/A
○	○	○	○	○	○	○	○

Professionalism - Relative (1-5)

Distinctly Below Peer Level 1	Somewhat Below Peer Level 2	At Peer level (Most residents should be at this level) 3	Somewhat Above Peer Level 4	Distinctly Above Peer Level 5	N/A
○	○	○	○	○	○

Comments

Remaining Characters: 5000

Interpersonal & Communication Skills - Absolute/Anchored (1-7)

Needed Significant Assistance 1	Needed Moderate Assistance 2	Needed only Minimal Assistance 3	Needed Very Infrequent Assistance 4	Performed in a Fully Independent Manner 5	Able to Serve as a Consultant to Other Physicians 6	Expert and Able to Serve as a Resource to Fully Trained Anesthesiologists 7	N/A
○	○	○	○	○	○	○	○

Interpersonal & Communication Skills - Relative (1-5)

Distinctly Below Peer Level 1	Somewhat Below Peer Level 2	At Peer level (Most residents should be at this level) 3	Somewhat Above Peer Level 4	Distinctly Above Peer Level 5	N/A
○	○	○	○	○	○

Comments

Remaining Characters: 5000

Systems-based practice - Absolute/Anchored (1-7)

Needed Significant Assistance 1	Needed Moderate Assistance 2	Needed only Minimal Assistance 3	Needed Very Infrequent Assistance 4	Performed in a Fully Independent Manner 5	Able to Serve as a Consultant to Other Physicians 6	Expert and Able to Serve as a Resource to Fully Trained Anesthesiologists 7	N/A
○	○	○	○	○	○	○	○

Systems-based practice - Relative (1-5)

Distinctly Below Peer Level 1	Somewhat Below Peer Level 2	At Peer level (Most residents should be at this level) 3	Somewhat Above Peer Level 4	Distinctly Above Peer Level 5	N/A
○	○	○	○	○	○

Comments

Remaining Characters: 5000

What are particular strengths of this resident?

Comments

Remaining Characters: 5000

Residents want to know how they can improve. Note specific areas or items that this resident should focus on improving:

Comments

Remaining Characters: 5000

Clinical Competency Evaluation:

This information is available to the Competency Committee. If you answer 'Yes' to any questions, PLEASE add a comment.

When working with this resident, I have concerns regarding patient safety:

Yes	No	N/A
○	○	○

I have concerns regarding this resident's honesty, ethics or character

Yes	No	N/A
○	○	○

I am concerned that this resident lags behind peers or may need extra help

Yes	No	N/A
○	○	○

I have concerns about this resident as a team player

Yes	No	N/A
○	○	○

I have concerns about this resident's openness to teaching and improvement

Yes	No	N/A
○	○	○

Comments

Remaining Characters: 5000

Do you have confidence that this resident can perform the following cases in an independent and unsupervised setting?

MAC for skin biopsy (healthy patient)
Yes No N/A
○ ○ ○

GA with an LMA for Knee arthroscopy (healthy patient)
Yes No N/A
○ ○ ○

Appendectomy (full stomach, otherwise healthy)
Yes No N/A
○ ○ ○

Ex Lap for perforated viscous (otherwise healthy)
Yes No N/A
○ ○ ○

Elective colectomy with coronary artery disease and chronic renal insufficiency
Yes No N/A
○ ○ ○

Elective craniotomy with increased Intracranial Pressure and asthma
Yes No N/A
○ ○ ○

Blunt trauma including head injury and liver injury
Yes No N/A
○ ○ ○

Leaking abdominal aortic aneurysm with decompensated congestive heart failure and acute atrial fibrillation
Yes No N/A

Have you reviewed this evaluation with the resident?
Yes No N/A
○ ○ ○

Please alert Program Director and Preceptor to this evaluation.
Yes No N/A
○ ○ ○